Examples of SAR-centric patent mining using open resources

Elsevier COMPREHENSIVE MEDICINAL CHEMISTRY III, July 2017, in press

Examples of SAR-centric patent mining using open resources

Christopher Southan, IUPHAR/BPS Guide to PHARMACOLOGY, Centre for Integrative Physiology, University of Edinburgh, EH8 9XD, UK.

Abstract

Structure activity relationships (SAR) published in journals underpin medicinal chemistry. However, patents contain more SAR data and surface years earlier. While their documents present challenges for data mining, there has been a recent "big bang" in the availability of extracted chemistry in open databases. Consequently, PubChem now contains ~20 million structures from patents, including most of those associated with bioactivity. This chapter covers a selection of resources, tools and tricks that can be used to dig out patent SAR. It also explores intersects between chemistry curated from papers by ChEMBL and automatically extracted from patents by SureChEMBL.

Keywords

Patents, papers, chemical structures, document text, SAR, databases, BindingDB, PubChem, SureChEMBL, ChEMBL, chemical images, IUPAC names, automated extraction, name-to-structure.

Introduction

This chapter will present examples of interrogating medicinal chemistry patent sources with the primary objective of digging out structure activity relationships (SAR). Since the details below will show this to be more difficult than for papers why bother? The main reasons are that patents not only contain more SAR but also that is published earlier. Statistics in support of both these advantages are difficult to obtain but will be partially addressed below. For discrete SAR values a conservative estimate of at least two-fold more data-mapped structures in patents can be made. If ranged values are included the difference rises to at least five-fold. In support of this, it was reported in 2011 that Factor IXa (F10, P00742) had 5.8K assayed compounds mapped to it via extracted from papers by ChEMBL but this had increased to 43K from combined literature and patent extractions by GVKBIO (i.e. papers: patents ~1: 6) [1]. The time advantage is more difficult to support with data but it is well known that pharmaceutical company drug discovery teams typically wait two to three years after a patent application before publishing in a journal. In some cases, high quality project results will be exemplified as substantial SAR sets in patents but may never surface elsewhere.

As an important adjunct to SAR extraction *per se* this chapter will also exemplify ways of exploring connectivity between papers and patents from the same teams. However, both aspects will be done using open databases and tools. Reasons for this restriction are outlined below;

- Commercial sources, even just major ones, now present too broad a spectrum of technical features to review (i.e. no author would have either the multiple licences or capacity for detailed comparative evaluation).

- There has been a remarkable "big bang" in the availability of automated patent-extracted chemistry just in the last few years, culminating in the secondary deposition of ~ 19 million patent-derived entries in PubChem and primary extraction of ~ 17 million of these into SureChEMBL [2]

- This has led to a democratisation of patent mining that has now become more accessible for those without commercial databases or tools.

- Concomitantly, parallel interrogation of open resources has become essential for users of commercial databases anyway, since the inevitable divergence of coverage by different sources precludes any one-stop-shop.

- PubChem precomputes relationships between 2D and 3D chemical structures, patent documents, PubMed IDs, PDB entries, bioassay results and annotation from many curated submitting sources [3]. This means that the combined PubChem and Entrez functionality now encompasses the majority of bioactive patent chemistry. This thus presents a scale of relationship navigation that closed commercial databases cannot match.

Although aspects of intellectual property (IP) focused searching and/or competitive intelligence (CI) analysis will only briefly be touched upon, strategically they may be better addressed by commercial databases, depending on what value-added features they offer for these two domains. This may include enhancing the indexing depth and precision of public data in various ways, including manual annotation.

Additional caveats related to the coverage of this chapter need mentioning. The first is that the techniques exemplified here are small-scale, largely manual and thus more individual orientated exercises rather than large scale approaches. The second coverage restriction is that illustrative examples can only be accompanied by short technical descriptions, although they will be referenced for users to access the details. The third is that the queries describe here we executed in September of 2016 and may thus give different results at later dates. The fourth is that, particularly in the patent domain, queries are rarely "clean" in terms of specificity or recall (some of the reasons for this will be discussed). The ones used here are presented in good faith with the expectation of some level of reproducibility. However, they can be disconcertingly noisy and inconsistent in practice. One reason is that patent offices and other sources often do not index document metadata in standardised ways (e.g. conflating applicants with inventors or multiple synonyms for the same applicant institutions). Another is that different interfaces to nominally the same document corpora can give different results (e.g. depending on the filtration options selected and the syntax used for complex queries). The fifth caveat is that while queries executed here have been designed to reflect real-world tasks they are also somewhat contrived for illustrative purposes and thus may not be the most efficient way to get to the explicit results. There are many choices of entry points, tools, query strategies and technical tricks that may answer particular questions more quickly.

Five final points related to the usability of this chapter should be mentioned.

1. With the exception of a few novel structures not yet submitted to any public databases specifically mentioned compounds will be specified either by a PubChem Compound Identifier or a Substance Identifier or in some cases both.

2. Rather than necessitate the addition of many hyperlinks into the text, each of these can be accessed via two standard URLs. For example, atorvastatin, as CID 60823 can be linked out as https://pubchem.ncbi.nlm.nih.gov/compound/60823. The specific submission from SureChEMBL, SID 226395935, can be analogously linked as https://pubchem.ncbi.nlm.nih.gov/substance/226395935.

3. There are too many patents to be fully referenced so their numbers are specified in the text. These will usually retrieve links to the document in the upper ranks of a Google search result within seconds.

4. SAR retrievals will focus on beta-secretase 1 (BACE1, Swiss-Prot P56817) as an Alzheimer's disease (AD) protease [4]. The illustrative advantages of an intensively pursued drug target will become apparent as we track through over a decade of substantial SAR and explore patent-to-patent and patent-to-paper connections.

5. For reasons of space, a basic familiarity with querying databases and patents has to be assumed. In particular this should include patent office websites (e.g. EPO Espacenet, USPTO and WIPO), kind codes (A1, B2 etc.) patent classifications, the concept of patent family and experience with searching in PubChem or other chemistry portals.

Statistics of extractable SAR

The complex task of extracting chemical and biological entities from patents has been well studied. Some of the different approaches will be referred to below but the complication is that SAR is defined by relationships between entities. For this reason, questions regarding SAR statistics both inside and outside patents are difficult to get precise answers to but would be useful for making judgments on what is realistically achievable.

In this context it is useful to review our high level assumptions. The first is that readers have an interest in bioactive chemistry in the wider sense, with the obvious proviso that very low or inactivity is crucial for SAR and control experiments but data sets are biased towards activity and potent compounds. The second would be that, regardless of the dominance of medicinal chemistry, IP domains related to tropical diseases, antivirals, antibiotics herbicides, pesticides, toxicology and chemical biology can be approached for SAR mining in broadly the same ways as described here for human target-centric drug discovery. It is also a reasonable assumption that medicinal chemists have at least some interest in extracted structures that were not designed to have biological effects *per se* (e.g. novel reactions, dyestuffs, photochemistry etc) but where the documented synthetic chemical space lies within the broad property envelope associated with potential bioactivity.

So what do we mean by SAR in the context of this chapter? The concept of a series of close structural analogues with a range of quantitative bioactivity data that can form the basis for different types of predictive modelling is well understood but also continually developing [5].

For *in vitro* data they need specific links between the entities of compounds, the protein bioentities whose activities they modulate and the assays used to measure this. These can be classified as (data supported) compound-to-assay-to-protein relationships [6]. Typically, a document "D" describes a biochemical assay "A" with a quantitative result "R" (e.g. an IC50) for compound "C" (with an explicit structure) defining it as an inhibitor of protein "P" (with a sequence and species-specific identifier). Useful shorthand for these relationships is thus "D-A-R-C-P" (note the substitution of "P" with target "T" can be used to indicate a black-box mechanism of action such as a tumour cell or a tropical disease parasite). This shorthand can be extended to describing a basic SAR series as a multiplexing of R and C, while a specificity cross-screen would be a multiplexing of A, R and P. Separate patent filings of SAR interest often cover closely related but distinct chemical series. These can be linked as a set of "D"s that have A-R-P in common (since they were from the same institution and reasonably close in time) but the different sets of "C" are varied.

Medicinal chemists have traditionally manually extracted the key entities from patents and papers of interest and organised the relationships locally, thereby converting unstructured data from document text, chemical images, result tables and supplementary data into structured information (e.g. D-A-R-C-P in an Excel sheet). It should also be pointed out that, beyond a sketching tool, access to patent office portals along with a basic familiarity of enzymology, receptor pharmacology, and protein naming, a scientist can curate A-R-C-P sets from a patent on the day of publication within a few hours (i.e. in advance of any commercial extraction service). Notwithstanding, this presupposes that a) the indexing of the title and abstract allowed the specific retrieval of patents of interest and b) that these actually contained novel, useful and extractable SAR.

While the ability to perform bespoke curation and annotation remains a core competence for medicinal chemists, they are increasingly likely to hand-off this activity for two main reasons. The first is that commercial or public databases, as well as internal support from Information Scientists, can provide project teams with partial or complete D-A-R-C-P extractions from various sources. For example, those with SciFinder access can collate the D-C records within days of publication but will still have to fill in A-R-C-P themselves (or from another resource). The second reason is that drug discovery enterprises (academic or commercial) usually have local systems for collating, mining and modelling SAR originating from their own internal projects. These are now likely to include the use of Electronic Laboratory Notebooks (ELNs) for primary chemical structure registration and data capture. These are then extracted to populate in-house databases that increasingly also integrate external bioactivity data sources. Pioneering examples of these include the AstraZeneca enterprise applications of Chemistry Connect [7] and SAR Connect [8]. What this means is that bespoke extractions are now expected to be format-compatible with internal systems and thus easily subsumed into them for intra-organisation sharing rather than languishing in local directories.

An expanding range of curated compilations containing partial or complete D-A-R-C-(P or T) mappings are available as commercial products and deposited in public databases [9]. Depending on scale and update frequencies, these largely obviate the need for individuals to "roll their own" extractions. However, a number of key points need to be made before reviewing some of these in the context of this chapter. Firstly, the facility with which a medicinal chemist can extract a document is reflected in the resource curation model (commercial or academic) where professional biocurators are hired or contracted for scaling-up essentially the same process. Secondly, full D-A-R-C-P extraction with useable specificity

has so far defied complete automation. While text mining combined with Natural Langue Processing (NLP) is making progress towards this goal, challenges remain. The principle one is the need to semantically link structures, assay descriptions, result tables and protein identifiers extracted from different sections of a document. Experienced biocurators can cognitively discern these relationships in a matter of minutes but they increasingly use automated tools for triage and to improve their efficiency of capture. Thirdly, while it actually violates the principle of patenting, some applicants in the medicinal chemistry domain deliberately introduce varying degrees of obfuscation into their documents (examples will be mentioned). In practice this can not only make the elucidation of SAR relationships more difficult but in some cases impossible.

The third point concerns patent coverage. As testimony to the difficulty of managing a patent curation pipeline, compared to processing papers, even now relatively few resources can surface complete D-A-R-C-P sets from patents. Some commercial offerings may index a molecular mechanism of action (mmoa) but this is only a partial substitute for a standard protein identifier. The fourth point is that, in many cases, it is necessary to compare sources indexed for bioactivity as a surrogate for SAR *per se*. Consider, for example, a published report on an orphan (i.e. with no analogue structural neighbours) complex natural product that potently inhibits malaria in a mouse model. A record of this bioactivity is clearly valuable and worth finding but is not immediately useable for SAR exploration. In contrast, extended data sets either where "T" uses growth inhibition of the parasite in vitro with 100s of IC50 values for a designed synthetic library and/or recording Ki's for the identical chemical series against one or even multiple "P"s as purified *Plasmodium* enzyme targets. While filtering database entries between these different bioactivity levels can sometimes be done for papers, this is more difficult for patents.

The content statistics of three databases in table 1 are partitioned into D-A-R-C-P metrics. The reason for selecting these in particular are not only that the relevant numbers are available from the teams concerned but also that, in combination, they sample the global "bioactivity space" and thus provide useful overviews of what could potentially be SAR-mined.

Table1. Document, assay, result, compound and protein target content counts for Excelra GOSTAR (commercial), BindingDB and ChEMBL22 as recorded in Sept 2016. Numbers for the latter were taken from the release notes except the protein target count was obtained from

the UniProt cross-reference. Those from BindingDB were personally communicated by Michael Gilson and the Excelra figures by Anil Kumar and Sreeni Devidas

Name	D (pubs)	D (pats)	A	R	C	P
GOSTAR	110K	69K	17 mill	24 mill	7.3 mill	9212
BindingDB	25K	1.2K	0.96 mill	1.3 mill	0.57 mill	6468
ChEMBL	65K	1.2K	0.9 mill	5.6 mill	1.2 mil	6888

There are caveats associated with table 1. These include that the individual data models give rise to differences in entity counts, redundancy reduction and relationship definitions. Note also that GOSTAR is a merge of a suite of databases. Nonetheless, table 1 provides a useful quantitative overview. We can briefly go through them in sequence. As a commercial database, GOSTAR is the largest manually curated resource of chemical modulators of biological targets (mostly proteins) with associated activity data (mainly inhibition) extracted from papers and patents. Notwithstanding the fact it cannot serve as an example of open patent mining, it nonetheless provides data-supported upper limits of extractable SAR from public documents. A 2013 analysis indicated the journal: patent compound split was approximately 1:2.7 with a document number split of 82K : 58K [10]. A recent update from Exelera (table 1) records current counts as 2.2 million structures from the literature, 5.1 million from patents (i.e. a slight shift towards papers since 2013) and an overlap of 0.19 million (i.e. not much change). Other relevant figures from the 2013 publication included that manual extraction yielded, on average, 12 compounds from a paper and 45 from a patent.

The BindingDB figures are derived from an open database, not only *in situ* but also as a regular submitter to PubChem [11]. In addition, this resource supplies a set of D-A-R-C-P expert manual extractions of US patents (reviewed later). Since 2009 ChEMBL, also open and a PubChem submitter, is an open source of literature-extracted bioactivity and it also subsumes confirmatory PubChem BioAssays [12]. Full release statistics are available from which the headline numbers in table 1 were obtained. Between BindingDB and ChEMBL there are similarities, differences and complementarity in terms of SAR content. Firstly BindingDB subsumes protein-mapped ChEMBL content. Consequently, the former has 78% of compounds in common with the latter (although a proportion of these include identical structures collated by different routes). In addition, BindingDB has extracted 3K PubMed IDs not in ChEMBL. The average of 18.5 extracted compounds per paper for ChEMBL22 is slightly higher than for BindingDB at16, but below the current GOSTAR average of 20. Thus, these three independent numbers indicate a reasonably consistent yield of manual compound extraction from the medicinal chemistry literature along with the linked A-R-P data.

So just how much total SAR is extractable from the journal corpus? This question provides context for comparisons with patents but it is not easy to get a clean answer. On balance, the

2.2 million compounds from papers extracted into GOSTAR indicate a plausible lower boundary, since the spread of journals and curation backfill time span is larger than ChEMBL. However, quantifying how much of this is useful for SAR modelling (e.g. a series of analogues against the same protein as opposed to singletons with just one assay result) is less straightforward (but the distribution could probably be discerned from GOSTAR internal statistics). Upper boundaries remain uncertain but what we can record is that 92K PubMed entries are classified under the MeSH term "Chemistry, Pharmaceutical" and 713K under "Chemistry". While GOSTAR would not be expected to cover all bioactive chemical structures ever published, by their own selection criteria of SAR-density per-paper, they have probably captured well over the major part. Given the PubMed figures above we can thus guestimate a conservative upper limit for ~3 million structures as being within the wider frame of SAR interest. Notwithstanding, as will be expanded on, ChEMBL constitutes the largest open resource of literature-extracted SAR against which patent sources and subsets can be intersected.

Statistics of putative patent SAR

As we have seen for the literature, precise comparative statistics for document corpora entity content are difficult to come by. Notwithstanding, useful calibrations from automated extractions with a 2015 time point have been provided by SureChEMBL and the Minesoft Chemical Explorer (commercial). The former declare 17 million compounds from 14 million patent documents while the latter claim 12 million from 10 million. While the gap between these numbers and the 5.1 million selected for SAR content from the same corpus for GOSTAR is wide, we can reconcile them to a certain extent via other methods of analysis and orthogonal queries to provide broadly plausible cuts. The first filter we need to consider is patent family redundancy. Using the World Intellectual Property Organization (WIPO) PATENTSCOPE database we can establish that 58 million national documents collapse to 3 million international patent applications (PCTs or WO filings). This indicates an average family size of 19 that, for the purposes of mining, can be considered as having the same chemistry content (despite post-examination claim scope changes). The other advantage of selecting WO is that they are generally the first family members to publically surface (for the record thought, these are technically only published applications since granting is done by the national authorities). Their disadvantage for mining is that, compared to the XML format of USPTO applications from 2001 onwards, WOs have poorer text and image quality from suppliers of the electronic transformations. This degrades chemistry extraction, sometimes severely (this topic is addressed later).

The second key filter is to identify relevant chemical content. As a first step towards the latter we can use the Intentional Patent Classification (IPC, see website listing). This is a hierarchical system where C for "chemistry" accounts for 20% of the WO. However, the combination of C07 for "organic compounds" and A61K "preparations for medical purposes", while not 100% specific, is highly selective for small molecule drug discovery, medicinal chemistry and therefore potential SAR content (n.b. not all relevant national filings proceed to a PCT and some not captured in this intersect may be using submission tricks to "hide"). The individual numbers, via PATENTSCOPE, are 217K and 235K respectively with a C07/A61K intersect of 85K (for comparison A01N "herbicides and insecticides" has 29K WO with 7.4K of these intersecting with C07). The corresponding C07/A61K count for EPO publications is 371K which drops to 233K for USPTO. Despite the uncertainties on

absolute document redundancy (e.g. further collapsing by Kind codes) we can make a plausible estimate that the medicinal chemistry relevant patent subset could be in the range of 100K to 200K. We can then use the average count of 257 unique IUPAC structures per patent from recent exhaustive extraction exercise to make a total chemistry estimate for the C07/A61K set of just over 25 million.

Compared to document counts in the high millions often highlighted by patent indexing and extraction resources, these numbers for the medicinal chemistry intersect seems lower than might be expected. However, it is comparable to the GOSTAR extracted patent count of ~ 70K in table 1 which is SAR selective. The yearly WO published output corresponding to C07/A61K can be plotted (figure 1). The result clearly mirrors the published rise and fall in the concomitant GOSTAR chemistry extraction indication the amount of patent SAR has actually declined since ~2006 [10].

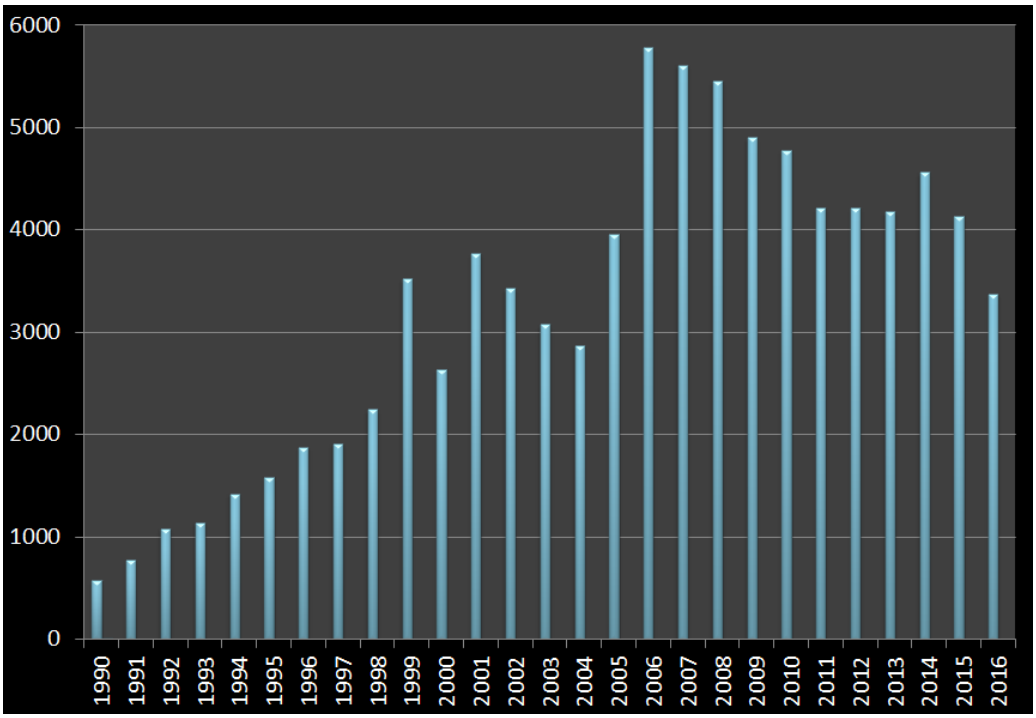


Figure 1. A plot of the per-year WO patent counts for the intersect between the IPC codes C07 and A61K. The September 2016 total was 85K.

While commercial operations are triaging this output, it is difficult to get open statistics that could support SAR estimates. For example, the family redundancy between the WO, USPTO and EPO is unclear (even though some measure might be obtainable from the INPADOC system). Note it is possible to select granted vs non granted applications in USPTO as a nominal quality filter and redundancy reduction. However, not only does the time lag for examination run into years but also granting is by no means a guarantee of scientifically useful content. What we can attempt is to empirically divide C07/A61K into data-centric utility groups (i.e. not based on IP considerations) with respect to putative SAR content. These can be outlined as follows;

1. An extended chemical series, including synthesis and analytical details with novel composition of matter, described as directed towards a specific target and a set of diseases, possibly including assay descriptions. However, beyond a general

statements, such as “found to be active”, no data can be assigned to individual structures.

2. As above but with a proportion of the examples having aligned bioactivity data, although much less than the number of example structures and/or heavily gapped.

3. Assignments of activity to each exemplified structure but only as ranged values. This can be typically in the form “between 10 and 100nM or 100 to 500 nM” or using a binning system of one to five stars, ABCDE or suchlike.

4. The document includes a complete set of discrete assay results, typically IC50 or Ki, displayed in a data table with anywhere between 5 and 500 rows. These can be aligned with the structures (novel at filing time) and their synthetic descriptions. By definition, this constitutes the most useful SAR patent (particularly if cross-screening data from two or more parologous proteins are included) but note the results may be generated with assay formats that do not use purified proteins (i.e. “T” not “P”)

5. Documents along the lines of 3 of 4, ostensibly including an extensive set of results, but where the degree of obfuscation in the description or layout (e.g. activity and example numbering not being in register) is high enough to preclude unequivocal SAR mapping.

6. A category that encompasses what we can term secondary filings in the sense that structures have been filed previously, are not novel and few in number. These can include crystallisation or synthetic route descriptions by the erstwhile originators of the lead compounds as well as generic manufacturers jockeying for inventive claims on clinical candidates.

7. Outside the above there are many filings that defy any simple classification. They are still assigned the two codes we deem useful but for various reasons difficult to discern SAR. Perhaps the most interesting are data supported authentic new uses for previously claimed structures (i.e. repurposing). Others may cover mixture combinations, often with huge permutated listings. Some include various types of virtual enumeration where potential activities for structures are claimed that have neither been made nor tested.

The informal categories above can be assigned by detailed inspection of patent documents. However, they expose the problem that there are no statistics as to how these are partitioned at scale. Consequently, there are no reliable methods of automated SAR detection. While 2,3, and 4 are of primary interest, some of the other categories could be useful. For example, in category 1 there are fillings by drug discovery teams with an established reputation but whose project leaders or company management take the strategic decision to file a chemical series without disclosing any activity data. Notwithstanding, these are logically expected to include actives (i.e. a series with cryptic SAR) to justify the IP investment in the first place. They can certainly be used for modelling (e.g. building a pharmacophore) and some structures may be selected as a smaller SAR series and published later in a journal.

Identifying specific patents

Retrieval of patents with a view to significantly enriching for categories 2,3, and 4 above, is a major subject that cannot be covered in depth here. However, examples will naturally emerge as we explore queries in different resources. In the interim, we can outline the general challenges but even before this we need to clarify operational imperatives in the context of this chapter. In searching databases that include patent-extracted chemistry in the context of IP due-diligence the consequences of false-negatives (e.g. overlooking relevant prior art) for any drug discovery organisation can be dire in terms of economic cost and erosion of competitive position. However, the assumption is made here that if an individual happened to miss some patent SAR that was retrospectively found to be relevant they would not be held to account too severely (although the same oversight from patent attorneys or Information Sciences teams might not be seen in such an equable light). Another real-world consideration is that researches are likely to be more focused on the SAR of the present rather than the past. This is especially so since the latter is well covered by review articles (e.g. the journal " Expert Opinion on Therapeutic Patents"). This means researches are unlikely to want to pull back SAR from decades ago and are thus more likely to filter for more recent time spans (where the chemotypes, potencies and properties should, at least implicitly, be improving) . Thus, in regard to the strategic dilemma of specificity vs recall, examples used here will be high-specificity queries and recent time scales in order to reduce time spent triaging false-positives for rejection, rather than being overly concerned about missing false-negatives or older filings.

As we know patents can be identified by their metadata such as inventor names, affiliations and patent classification codes (even though these can have noisy indexing). Filtering or ranking by publication date is of course useful but can be confounded by families, Kind codes and priority equivalents which can surface identical documents across time spans of five years or more. However, examples below will concentrate on target-centric searching that, while certainly prone to false-negatives, has an acceptable specificity. In addition, target searching is the most common first-pass for literature searching so that grappling with protein naming issues becomes more familiar.

Chemistry extraction

As mentioned, manual curation of SAR from a patent is intuitively obvious to medicinal chemists, many of whom have engaged in generating the data that goes into them and may even have had a hand in drafting applications. However, most sources that have pursued the large-scale manual extraction of chemical structures from patents (with or without A-R-C-P mapping) have been commercial. This means that technical details of their selection and triage are not usually disclosed. However, being proprietary about exactly how databases are populated always leaves users with uncertainties. In particular, it is important to know curation strategies and chemistry business rules that might directly affect how easily and how much SAR can be collated from any particular resource (as we will see below). These include

- How are ambiguous partial stereo centres handled?
- Is the parent added as an extra structure where the document specifies a salt?
- Are obvious applicant-originated chemical errors fixed?

- What activity units do they normalise to?
- Which software do they use to output their sketched structures?
- What primary standardised chemical representations are captured from this output?
- Are activity figures rounded down?
- Do they number structures corresponding to their example numbers in the document (i.e. so the extracted structure can be unequivocally matched to the activity values even if it is exemplified multiple times)?
- Do they have capping rules for very large SAR tables?
- Are all examples extracted or only key structures selected by the curators?
- Do they add the structures of known drugs or clinical compounds even though may these may only be specified by a semantic name?

Users who commit a lot of searching time to any particular curated resource may have to work these rules out "backwards" during the course of their analyses. They can also compare between sources (e.g. for the same patent number) but this does become time consuming.

By definition, the manual throughput of chemistry and other entity extractions (e.g. A-R-C-P) from documents factors linearly with the cost of biocurator time. For this reason, the technology with the greatest impact on open chemistry extraction in recent years (and provision of structures that can potentially be SAR-mapped) has been automated Chemical Named Entity Recognition (CNER). This scales approximately according to computing power and algorithmic efficiency. In essence, electronically formatted documents go into a pipeline and structures come out. As well as indexing the chemistry, outputs may be tagged with metadata, such as whether they were derived from text and/or images, recognised in the title, abstract, description or claim sections. In practice CNER extends beyond just text-to-structure (n2s) to encompass automated image-to-structure (i2s) as well as manually draw ChemDraw files then automatically converted to molfiles structures. These are generated by the USPTO Complex Work Units (CWU) process as part of their pipeline for complete electronic processing of patent applications. While researchers mining patents do not need to keep abreast of the details of CNER technology, it is important to understand the associated strength and weaknesses, particularly from the point of view of exploiting resources derived from it. A useful schematic describing SureChEMBL is shown in fig.3 (interrogation of this database will be described later).

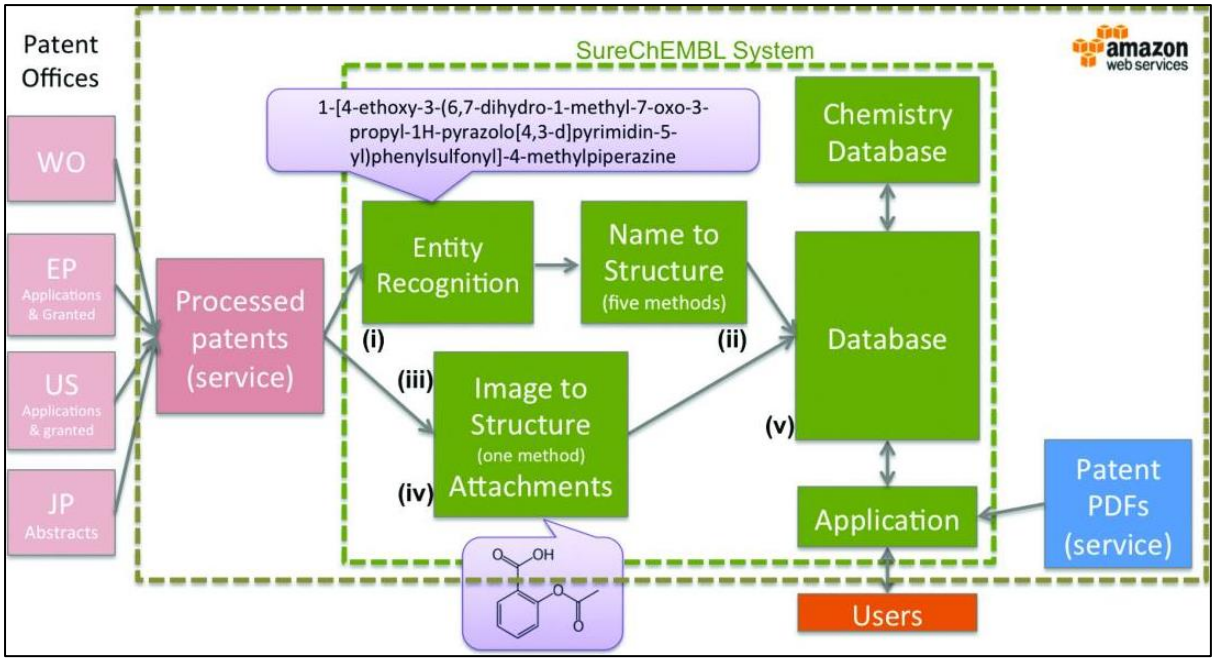


Figure 3. A diagram of the SureChEMBL CNER pipeline (adapted from [13]).

The basic operations shown in fig. 3 are n2s and i2s. The first is via conversion of IUPAC systematic names as a text strings but this also encompasses what we can call semantic n2s where names such as "atorvastatin" are converted into explicit structures via a look-up dictionary. The second is via image processing for i2s that also includes the CWU-derived molfiles. We can now move on to see how the interplay between manual inspection and automated extraction can be exploited for SAR retrieval.

Patent SAR retrieval from BindingDB

Before assessing CNER resources it is instructive to explore one of the few open resources that include manual patent extraction of complete D-A-R-C-P sets. As of Sept 2016 BindingDB had curated 1229 US Patents with target-directed small molecule binding data. While a relatively small number (now mirrored in ChEMBL22) it has the advantage of being produced by the expert triage of recent USPTO publications selected via the inclusion of Kd, IC50, Ki or EC50 in the abstract. The triage also selects for CWU molfiles that provide complete sets of example structures, thereby enabling the expert manual curation to focus on target identification and activity mapping. There is a pointer to this section from the front page or lower down on the left hand facets under citation sources. There is also an Advanced Search tools option for focused searches within just the patent data (or other sets). The top-ten patent extraction records by date added are shown in fig.3

Patent: (prefix with "US" like US6078908) Search

Patent	Data	Patent Title	Organization	Deposition
US9199947	140	Biaryl-propionic acid derivatives and their use as pharmaceuticals	Sanofi	09/20/16
US9199927	112	Guanidinobenzoic acid compound	Astellas Pharma Inc.	09/20/16
US9199944	155	N2,N4-bis(4-(piperazine-1-yl)phenyl)pirimidine-2,4-diamine derivative or pharmaceutically acceptable salt thereof, and composition containing same as active ingredient for preventing or treating cancer	KOREA RESEARCH INSTITUTE OF CHEMICAL TECHNOLOGY	09/20/16
US9199949	11	Selective glycosidase inhibitors and uses thereof	Alectos Therapeutics Inc.; Merck Sharp & Dohme Corp.	09/20/16
US9193689	18	3-aryl-5-substituted-isoquinolin-1-one compounds and their therapeutic use	INSTITUTE OF CANCER RESEARCH: ROYAL CANCER HOSPITAL (THE)	09/19/16
US9187487	46	Azaindole derivatives as tyrosine kinase inhibitors	PRINCIPIA BIOPHARMA, INC.	09/19/16
US9192603	67	Heterocyclic sulfone mGluR4 allosteric potentiators, compositions, and methods of treating neurological dysfunction	Vanderbilt University	09/19/16
US9193697	39	Oxazole derivatives useful as modulators of FAAH	MERCK SHARP & DOHME CORP.	09/19/16
US9193736	311	PDE 10a inhibitors for the treatment of type II diabetes	Janssen Pharmaceutica, NV	09/19/16
US9193762	57	Selective inhibitors of prolylcarboxypeptidase	UNIVERSITY OF MISSISSIPPI	09/19/16

Figure 3. Entries from the BindingDB patent extraction page. The numbers under the "Data" column are measurement counts. Each of these has a PubChem CID derived either from a BindingDB submission (as an SID) or it matches a pre-existing CID from another source. The majority are unique within this patent set but there is some multiplexing for dual values (e.g. Ki and EC50) or multiple target cross-screens.

These BindingDB true-positives (i.e. most being from category 4 above) serves not only as a useful introduction the principles of SAR mining but also to benchmark other resources. A disadvantage is that, as USPTO publications, these can lag some time behind the surfacing of WOs. For example the most recent in the list, US9199947 published on Dec 1, 2015, was first published as WO2014154727 on Feb 2nd, 2014. However, since it is the availably of the CWU structures that make this pipeline possible in the first place, WO documents cannot be used. As a first look at the set, we can plot the number of compounds per document to display the SAR distribution (fig. 3).

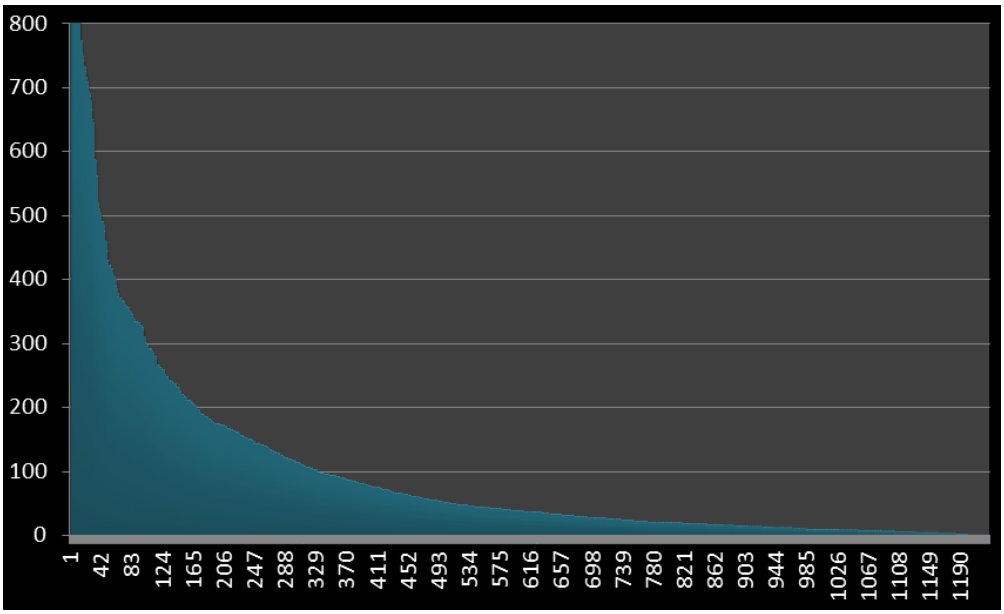


Figure 4. 1218 curated patent documents from BindingDB (horizontal axis) ranked by number of compounds x activities (the 14 with over 800 were truncated out of the vertical axis range). Over 80% have 20 or more SAR records and nearly 30% include at least 100. (n.b. the average measurement record per-patent of 98 is comparable to the GOSTAR non-redundant count of 65 structures per patent).

The most common first-pass strategy is to search by target name. However, as we can see in the list, the use of unambiguous protein names by applicants is patchy and titles such as "Chemical compounds" (US9156831) immediately present the false-negative problem. Notwithstanding, in this case cell proliferative disorders is mentioned in the abstract, the description goes on to specify PI3K-α and PI3K-δ and BindingDB curation has resolved the activities against PIK3CA, P42336. We can move on to the question "how many BACE1 patents can we find in this list?". Note that the BindingDB has already resolved the targets during their triage (i.e. there are 8030 records linked to BACE1, mostly from papers) but we in this case we can use the text query to illustrate general principles of patent retrieval.

Table 2. Entries from BindingDB patents that can be retrieved by related target names. The first 14 rows are matches for "BACE", ranked by compound count. The last two rows are matches to "memapsin" and "beta secretase".

Patent	Data	Patent Title	Organization	Deposition
US9181236	290	2-spiro-substituted iminothiazines and their mono- and dioxides as bace inhibitors, compositions and their use	Merck Sharp & Dohme Corp.	08/29/16
US8865911	180	Compounds and their use as BACE inhibitors	Astrazeneca AB	04/06/2015
US8754075	87	1,3-oxazines as BACE1 and/or BACE2 inhibitors	Hoffmann-La Roche Inc.	11/11/2014
US8748418	53	1,4-oxazepines as BACE1 and/or BACE2 inhibitors	Hoffmann-La Roche Inc.	11/11/2014
US8541427	45	Phenyl-substituted 2-imino-3-methyl pyrrolo pyrimidinone compounds as BACE-1 inhibitors, compositions, and their use	Merck, Sharp & Dohme, Corp.	11/26/13
US9000182	41	2H-imidazol-4-amine compounds and their use as BACE inhibitors	AstraZeneca AB	10/19/15
US9145426	38	Pyrrolidine-fused thiadiazine dioxide compounds as BACE inhibitors, compositions, and their use	Merck Sharp & Dohme Corp.	06/13/16
US9000184	31	Cyclohexane-1,2′-naphthalene-1′,2″-imidazol compounds and their use as BACE inhibitors	AstraZeneca AB	10/19/15
US9067924	24	1,4 thiazepines/sulfones as BACE1 and/or BACE2 inhibitors	HOFFMANN-LA ROCHE INC.	02/22/16
US9000183	21	Cyclohexane-1,2′-indene-1′,2″-imidazol compounds and their use as BACE inhibitors	Astrazeneca AB	10/19/15

US8940748	21	Iminothiadiazine dioxide compounds as BACE inhibitors, compositions, and their use	Merck Sharp & Dohme Corp.	06/16/15
US9000185	18	Cycloalkyl ether compounds and their use as BACE inhibitors	AstraZeneca AB	10/19/15
US8703785	9	2-aminopyrimidin-4-one and 2-aminopyridine derivatives both having BACE1-inhibiting activity	Shionogi & Co., Ltd.	10/14/14
US9029367	5	BACE inhibitors	Eli Lilly and Company	11/16/15
Synonym matches				
US9096541	28	Inhibition of memapsin 1 cleavage in the treatment of diabetes	Oklahoma Medical Research Foundation; Purdue Research Foundation	03/28/16
US8637504	37	Sulfur-containing heterocyclic derivative having beta secretase inhibitory activity	Shionogi & Co., Ltd.	08/22/14

The results in table 2 indicate some of the obstacles to target name resolution. An initial test with "BACE1" gave only four matches. A stemming variation "BACE" gave 17 matches but only to the 14 records in table 2. It happens that BACE is both a synonym for BACE1 and a common stem for the paralogous pair of BACE1 and BACE2. Importantly for SAR aspects, these two are both targets for Alzheimer's Disease and type 2 diabetes, respectively [4] and are also sometimes used in reciprocal cross-screening (thus expanding the SAR). Hence "BACE" not only conveniently picks up all the BACE1 single-target filings but also US8754075 and US8748418 as "and/or" double-target filings [14] . We also find US8541427 that uses an incorrectly hyphenated BACE-1 as symbol. However, "BACE" misses two false-negatives that, from testing the common synonyms, we can retrieve with "beta secretase" and "memapsin 1". The latter is a synonym for BACE2 which makes US9096541 nominally the only BACE2 single-target filing in this set, but the data also include BACE1 cross-screening results. With the proviso of needing a registration log-in, BindingDB facilitates the full download of SAR sets. An example is shown in table 3.

Table 3. Selected structure-to-activity records from US8748418. The complete table of 53 rows and a detailed set of columns was edited down to just the columns and rows below and ranked by IC50.

BindingDB MonomerID	BindingDB Ligand Name	Target Name Assigned by Curator or DataSource	IC50 (nM)	PubChem CID
123867	US8748418, 12	Beta-secretase 2 (BACE2)	5	68307545
123862	US8748418, 6	Beta-secretase 2 (BACE2)	11	60194542
123857	US8748418, 1	Beta-secretase 2 (BACE2)	24	66547140
123869	US8748418, 14	Beta-secretase 2 (BACE2)	28	66547488
123864	US8748418, 8	Beta-secretase 2 (BACE2)	30	66547317
123862	US8748418, 6	Beta-secretase 1	30	60194542

123858	US8748418, 2	Beta-secretase 2 (BACE2)	35	66547141
123858	US8748418, 2	Beta-secretase 1	40	66547141
123870	US8748418, 15	Beta-secretase 1	40	66547492
123870	US8748418, 15	Beta-secretase 2 (BACE2)	58	66547492
123857	US8748418, 1	Beta-secretase 1	70	66547140

Importantly, these can be related to their positions in the patent document and detailed connectivity information is available in the BindingDB entry (figs. 5 and 6).

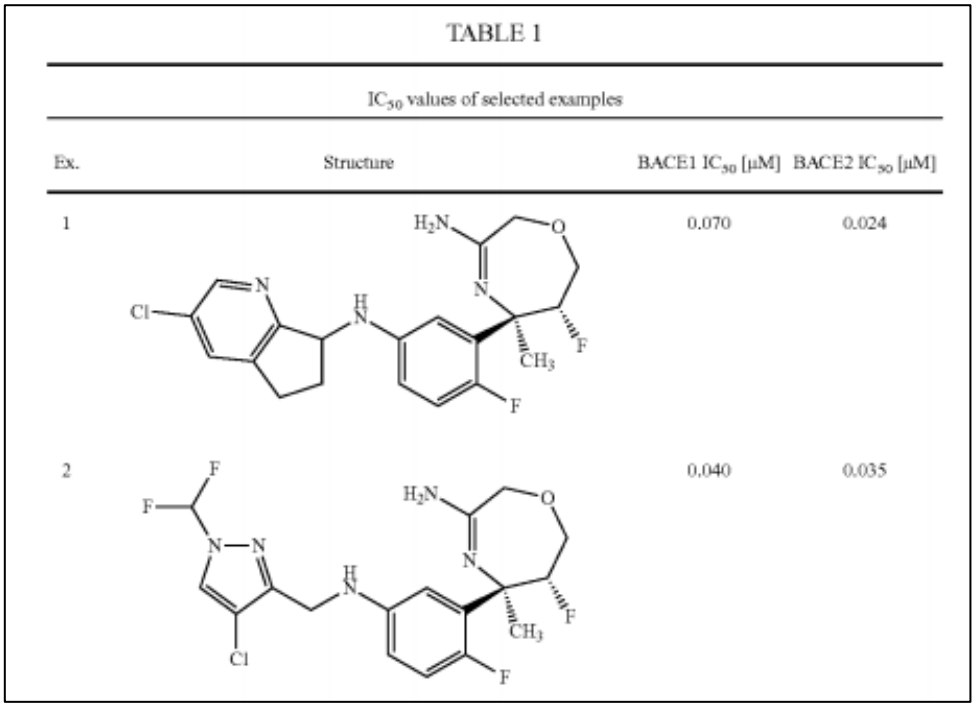
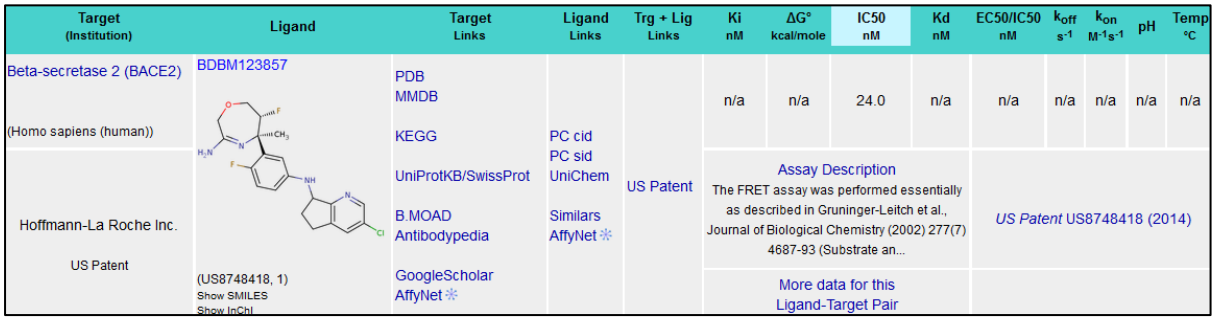


TABLE 1

IC_{50} values of selected examples

Ex.	Structure	BACE1 IC_{50} [μM]	BACE2 IC_{50} [μM]
1		0.070	0.024
2		0.040	0.035

Figure 5. Original SAR Data from US8748418 aligned with the example images and IC50 data from page 32 of the PDF. Example 1 = BindingDB 123857=US8748418, 1 = CID 66547140. For Example 2 = BindingDB 123858 = US8748418, 2 = CID 66547141



Target (Institution)	Ligand	Target Links	Ligand Links	Trg + Lig Links	Ki nM	ΔG° kcal/mole	IC50 nM	Kd nM	EC50/IC50 nM	k_{off} s^{-1}	k_{on} $M^{-1}s^{-1}$	pH	Temp °C
Beta-secretase 2 (BACE2) (Homo sapiens (human))	BDBM123857	PDB MMDB KEGG UniProtKB/SwissProt B.MOAD Antibodypedia GoogleScholar AffyNet	PC cid PC sid UniChem Similars AffyNet	US Patent	n/a	n/a	24.0	n/a	n/a	n/a	n/a	n/a	n/a
Hoffmann-La Roche Inc. US Patent	(US8748418, 1) Show SMILES Show InChI				Assay Description The FRET assay was performed essentially as described in Gruninger-Leitch et al., Journal of Biological Chemistry (2002) 277(7) 4687-93 (Substrate an...				*US Patent* US8748418 (2014)				
					More data for this Ligand-Target Pair								

Figure 6. The database entry for BindingDB 123857. The IC50 is specified along with with a citation for the assay. In addition many outlinks are presented such as the protein target in UniProt, the patent document (in Google and the USPTO portal) and the structure in PubChem (PC cid = CID 66547140)

The utility of this "pre-cooked" SAR is clear. However, if significant downstream effort is going to be invested in the data set (e.g. modelling and/or synthesis) it is prudent to cross check extractions (from any source in fact) against the original document. In this case US8748418 reveals 28 examples in the patent table 1, thus indicating the 53 BindingDB records have both BACE1 and BACE2 assay results as the SAR set. We can now compare this with a CNER source.

Patent chemistry in SureChEMBL

In a relatively short space of time SureChEMBL [13] has become the major public source of chemistry from patents since it had accumulated 17.25 million structures by the end of August 2016 (as indexed in UniChem [15]) . This not only offers advantages of scale but also speed because *in situ* (i.e. without the lag time associate with feeding-on to other databases such as UniChem or PubChem) chemistry from an individual patent document is downloadable within days of publication. The main limitation is that the primary relationship capture is only D-C. This means that users have to connect A-R-P data for the collation of SAR. As an example we can compare what SureChEMBL extracted from US8748418 where we have just located the SAR series manually extracted by BindingDB. One of the many routes to query SureChEMBL is shown in figure 7.

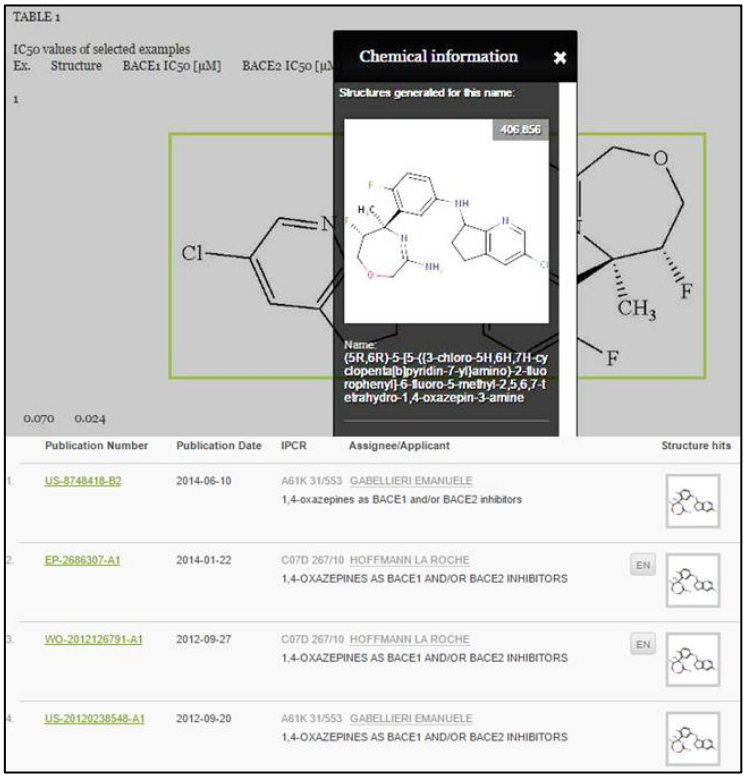


Figure 7. SureChEMBL queries related to the BindingDB 123857 structure. The upper panel shows the result of an exact-match structure search. This gives 9 intra-document matches, one of which an image-conversion (corresponding to figure 5) with the IC50 values in the lower left. The lower panel shows the inter-document matches across four family members.

We can download a table of the entire (i.e. unfiltered and redundant counts) chemistry records that SureChEMBL has extracted from US8748418 by a combination of n2s, i2s and CWU molfiles. This gives 876 rows. Extending this to all the family members increases to 1581. We thus see the main advantage of CNER in presenting what is probably close to total chemistry extraction. The important proviso is that achieving this is dependent on pre-processed patent text and image quality. The problem is that *a priori* from the download we cannot directly establish which of these have any SAR (i.e. could correspond to the 28 examples that BindingDB has already identified). Notwithstanding, this can be done manually within SureChEMBL by scrolling though the automatic mark-up of Table 1 in the document (i.e. each of the 28 images in this case). Insights into CNER output can be discerned in table 3.

Table 3. The 11 records in the SureChEMBL download for SCHEMBL12273617 (= BindingDB 123857 = CID 66547140)

patent_id	annotation_reference	schembl_id	type	doc_count	corpus_count
US-20120238548-A1	[3-((5R,6R)-3-Amino-6-fluoro-5-methyl-2,5,6	SCHEMBL12273617	TEXT	8	17
US-20120238548-A1	(5R,6R)-5-(5-(3-chloro-6,7-dihydro-5H-cyclop	SCHEMBL12273617	TEXT	8	17
US-20120238548-A1	US20120238548A1-20120920-C00015.TIF	SCHEMBL12273617	IMAGE	8	17
US-8748418-B2	[3-((5R,6R)-3-Amino-6-fluoro-5-methyl-2,5,6	SCHEMBL12273617	TEXT	8	17
US-8748418-B2	(5R,6R)-5-(5-(3-chloro-6,7-dihydro-5H-cyclop	SCHEMBL12273617	TEXT	8	17
US-8748418-B2	US08748418-20140610-C00015.TIF	SCHEMBL12273617	IMAGE	8	17
EP-2686307-A1	[3-((5R,6R)-3-Amino-6-fluoro-5-methyl-2,5,6	SCHEMBL12273617	TEXT	7	17
EP-2686307-A1	(5R,6R)-5-(5-(3-chloro-6,7-dihydro-5H-cyclop	SCHEMBL12273617	TEXT	7	17
WO-2012126791-A1	[3-((5R,6R)-3-Amino-6-fluoro-5-methyl-2,5,6	SCHEMBL12273617	TEXT	7	17
WO-2012126791-A1	(5R,6R)-5-(5-(3-chloro-6,7-dihydro-5H-cyclop	SCHEMBL12273617	TEXT	7	17
WO-2012126791-A1	imgf000037_0001.tif	SCHEMBL12273617	IMAGE	7	17

The key points are ;

- Multiple extractions of the same structure have been made from the four family members
- The n2s and i2s are concordant (i.e. alternative extractions produce the same ID)
- For the US application (A1) and grant (B2) the intra-document count was eight, split between two different IUPACs and one image
- The EP and WO missed one conversion (probably an image)
- The "corpus count" across all of SureChEMBL provides a reassuring inter-document consensus for this particular structure
- Manual inspection of the documents confirms the statistics of the data download in that Roche have exemplified their active series eight times
- This seems unusually high but we have no overall statistics to test this

The useful aspect of this example is that, using foreknowledge of the BindingDB extraction (albeit done at a later date) we can explore multiple ways of discerning the correct SAR series. One of these is via the corpus count shown in figure 8.

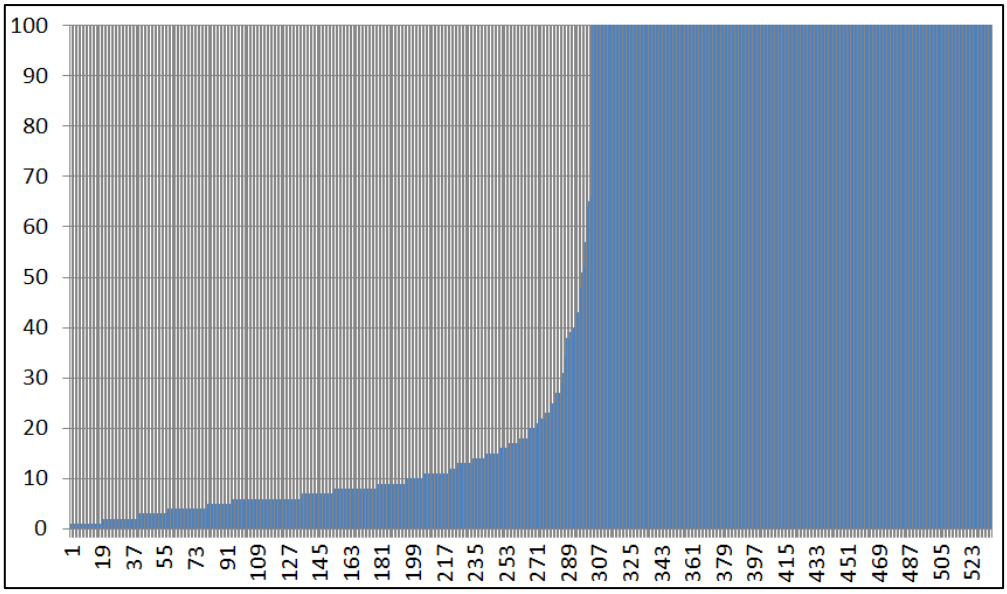


Figure 8. A plot of the corpus count (the vertical axis, truncated at 100 occurrences) for the non-redundant SureChEMBL chemistry download for US8748418.

The corpus count distribution of US8748418 shows the pros and cons of CNER pipelines in general. The problem is that most of the processing effort goes in to the continual re-extraction of frequently specified compounds (e.g. methanol occurs 7.2 million times across the SureChEMBL corpus and the name match finds 0.9 million documents). This is not to say that the statistics of common chemistry do not have utility for certain applications (e.g. surveying synthetic trends [16]) but these extractions can swamp out the bioactives. However, the leftmost section of fig.8 indicates that lower corpus counts probably represent a) novel structures and b) encompass the SAR series. However, while from table 3 we might expect 17 to be an SAR-specific cut-off, it turns out that the occurrences from 10 to 20 form a continuum of approximately 70 structures rather than a clean set However, a key feature of SureChEMBL is the ability to filter downloads as shown in fig. 9.

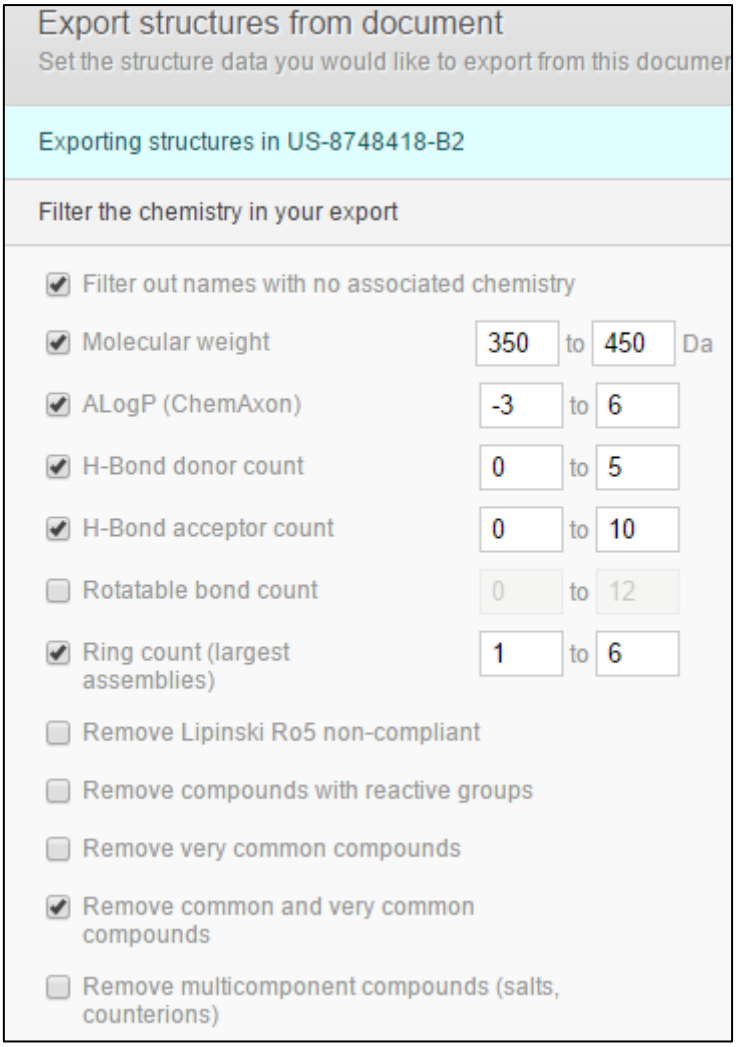


Figure 9. SureChEMBL chemistry export filters for individual patent documents or whole family extractions. Note the default Mw window is 300-700 but this was narrowed down since we have established that SCHEMBL12273617 (CID 66547140) has a Mw of 407.

Toggling these settings can be complemented by further post-download filtering (e.g. removing SureChEMBL ID duplicates and selecting occurrence ranges in the description section) to enrich for an SAR series. However, even with these options we are still left with 96 structures (i.e. within which should be the 28 we would like to identify). The caveat with Mw filtering is that it may encompass larger intermediates that are still novel structures but for which no bioactivity was reported. Corpus counting has the caveat that low frequencies of one or two (i.e. well below the document family count) may be chemically plausible structures but could represent artefactual extractions.

Regardless of the challenges of isolating an explicit SAR-linkable set, the facility to download patent chemistry extractions in SureChEMBL is powerful. There are however complementary ways to approach the isolation problem that can be revealed by manual inspection of context within the document. We have already seen this with the straightforward identification of the image structures and IC50 data in table 1 of the patent. A second useful option is to simply browse the SureChEMBL mark-up and spot sets of structures that look as though they could be matched to activity tables. An example of this is shown below for US8748418.

One embodiment of the invention is a compound of formula I, selected from the group consisting of

- [3-((5R,6R)-3-Amino-6-fluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenyl]-(3-chloro-6,7-dihydro-5H-[1]pyrindin-7-yl)-amine,
- (5R,6R)-5-{5-[(4-Chloro-1-difluoromethyl-1H-pyrazol-3-ylmethyl)-amino]-2-fluoro-phenyl}-6-fluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (5R,6R)-6-Fluoro-5-[2-fluoro-5-(tetrahydro-furan-3-ylamino)-phenyl]-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (5R,6R)-6-Fluoro-5-[2-fluoro-5-(tetrahydro-pyran-3-ylamino)-phenyl]-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (5R,6R)-6-Fluoro-5-[2-fluoro-5-(tetrahydro-pyran-4-ylamino)-phenyl]-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-5-[5-(6-Chloro-2,3-dihydro-benzofuran-3-ylamino)-2-fluoro-phenyl]-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-5-[5-(6-Chloro-pyridin-3-ylamino)-2-fluoro-phenyl]-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-5-{5-[(4-Chloro-1-difluoromethyl-1H-pyrazol-3-ylmethyl)-amino]-2-fluoro-phenyl}-6,6-difluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-5-{5-[(4-Chloro-1-difluoromethyl-1H-pyrazol-3-ylmethyl)-amino]-2-fluoro-phenyl}-5-ethyl-6,6-difluoro-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(1-methyl-1H-pyrazol-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(4-fluoro-phenylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(5-fluoro-pyridin-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(5-trifluoromethyl-pyridin-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(6-methoxy-pyridin-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(6-methyl-pyridin-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-6,6-Difluoro-5-[2-fluoro-5-(pyridin-3-ylamino)-phenyl]-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine,
- (R)-7-[3-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenylamino]-6,7-dihydro-5H-[1]pyrindine-3-carbonitrile,
- (R)-7-[5-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-2,4-difluoro-phenylamino]-6,7-dihydro-5H-[1]pyrindine-3-carbonitrile,
- (S)-7-[3-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenylamino]-6,7-dihydro-5H-[1]pyrindine-3-carbonitrile,
- [3-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenyl]-(3-chloro-6,7-dihydro-5H-[1]pyrindin-7-yl)-amine,
- [3-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenyl]-(6-chloro-2,3-dihydro-furo[3,2-b]pyridin-3-yl)-amine,
- [3-((R)-3-Amino-6,6-difluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenyl]-((S)-3-chloro-6,7-dihydro-5H-[1]pyrindin-7-yl)-amine,
- [3-((R)-3-Amino-6,6-difluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenyl]-((R)-3-chloro-6,7-dihydro-5H-[1]pyrindin-7-yl)-amine,
- 4-[3-((R)-3-Amino-6,6-difluoro-5-methyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenylamino]-benzonitrile,
- 8-[3-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-4-fluoro-phenylamino]-5,6,7,8-tetrahydro-quinoline-3-carbonitrile,
- (S)-7-[5-((R)-3-Amino-6,6-difluoro-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-5-yl)-2,4-difluoro-phenylamino]-6,7-dihydro-5H-[1]pyrindine-3-carbonitrile, and
- (R)-6,6-Difluoro-5-{2-fluoro-5-[6-(2,2,2-trifluoro-ethoxy)-pyridin-3-ylamino]-phenyl}-5,7,7-trimethyl-2,5,6,7-tetrahydro-[1,4]oxazepin-3-ylamine, or a pharmaceutical acceptable salt thereof.

Figure 10. List of IUPAC names from page 11 of the US8748418 PDF. The unbroken blue highlighting means that SureChEMBL has made a successful IPUAC conversion to a structure in each of the 27 cases. They would thus be included in the download. In addition each one has a link to information about the structure (as shown in fig. 7).

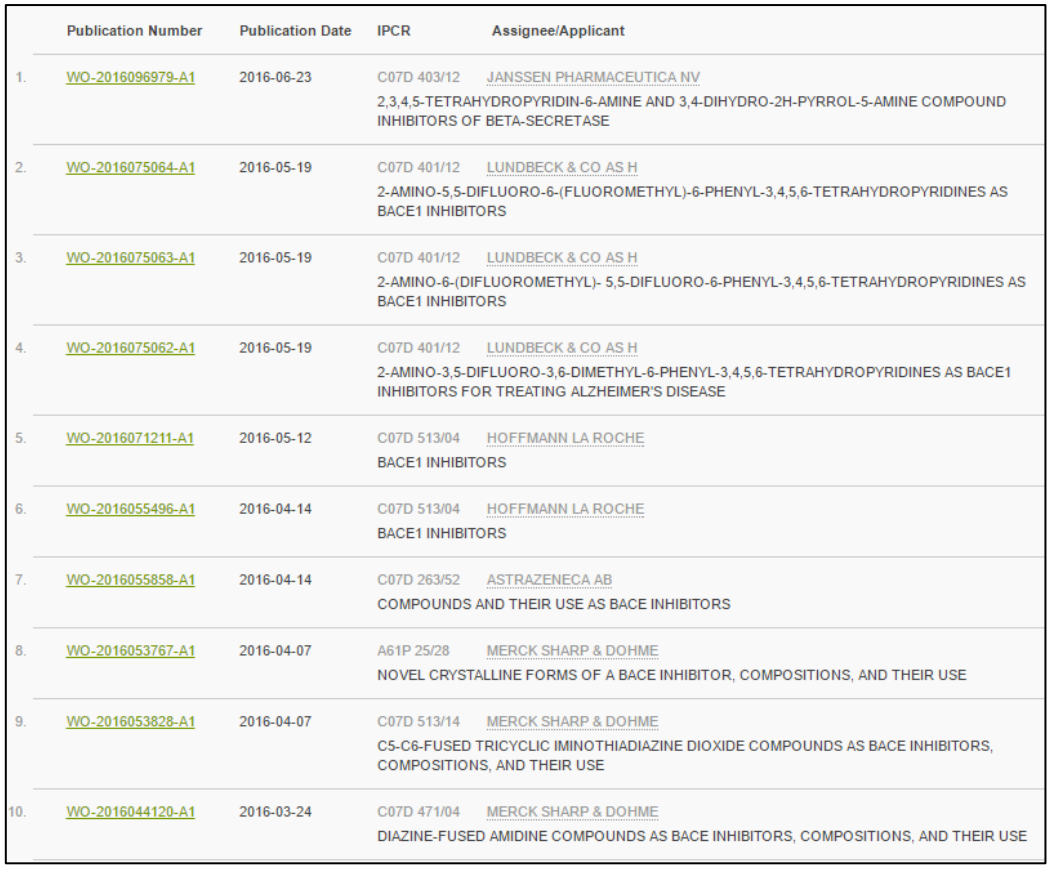
The ability to identify such large sections in patent mark-up (either as a blue highlighted IUPAC strings and/or a green border around the images) is useful for both for assessing the quality of extraction and delineating examples. In this case the series is not numbered so we cannot reliably link the structures to the data table rows in any automated way. However, example assignments (e.g. Example 1,2,3, etc.) can be found further down in the synthesis descriptions and thus manually collated via the compound displays. This example thus shows both the generic problem for SAR-digging and the solution via manual collation. Additional ways of establishing these structure-to-data links will be shown below.

Target searching in SureChEMBL

Having looked at target-specific retrieval in BindingDB we can compare this to searching the much larger patent set in SureChEMBL. It is also important to understand the major difference between these resources. The latter has indexed the full content of the docments whereas BindingDB only indexes what has been expert selected (but this includes the target name, UniProt ID, quantitative affinity values, etc.). We can look at a configured query in the SureChEMBL interface (fig.11)

Figure 11. A BACE1target query in the SureChEMBL interface showing four filters, query string (main box) location (doc sections toggle) patent classification code (lower panel) and WO as patent authority (right hand panel)

The filters in fig. 11 were selected primarily for specificity rather than recall (i.e. to give a good yield of true positives). The first filter is the HGNC gene symbol and two common protein synonyms as described for the BindingDB search in table 2. We have used C07D as the second select but we can skip A61K in this case since we are already selecting a biological target. The third select is to search the only the title and abstract rather than the entire text that can could include false-positives as incidental mentions of the target. The fourth select is for WO as already described. The retrieval results are shown in fig.12.



	Publication Number	Publication Date	IPCR	Assignee/Applicant
1.	WO-2016096979-A1	2016-06-23	C07D 403/12	JANSSEN PHARMACEUTICA NV
			2,3,4,5-TETRAHYDROPYRIDIN-6-AMINE AND 3,4-DIHYDRO-2H-PYRROL-5-AMINE COMPOUND INHIBITORS OF BETA-SECRETASE	
2.	WO-2016075064-A1	2016-05-19	C07D 401/12	LUNDBECK & CO AS H
			2-AMINO-5,5-DIFLUORO-6-(FLUOROMETHYL)-6-PHENYL-3,4,5,6-TETRAHYDROPYRIDINES AS BACE1 INHIBITORS	
3.	WO-2016075063-A1	2016-05-19	C07D 401/12	LUNDBECK & CO AS H
			2-AMINO-6-(DIFLUOROMETHYL)- 5,5-DIFLUORO-6-PHENYL-3,4,5,6-TETRAHYDROPYRIDINES AS BACE1 INHIBITORS	
4.	WO-2016075062-A1	2016-05-19	C07D 401/12	LUNDBECK & CO AS H
			2-AMINO-3,5-DIFLUORO-3,6-DIMETHYL-6-PHENYL-3,4,5,6-TETRAHYDROPYRIDINES AS BACE1 INHIBITORS FOR TREATING ALZHEIMER'S DISEASE	
5.	WO-2016071211-A1	2016-05-12	C07D 513/04	HOFFMANN LA ROCHE
			BACE1 INHIBITORS	
6.	WO-2016055496-A1	2016-04-14	C07D 513/04	HOFFMANN LA ROCHE
			BACE1 INHIBITORS	
7.	WO-2016055858-A1	2016-04-14	C07D 263/52	ASTRAZENECA AB
			COMPOUNDS AND THEIR USE AS BACE INHIBITORS	
8.	WO-2016053767-A1	2016-04-07	A61P 25/28	MERCK SHARP & DOHME
			NOVEL CRYSTALLINE FORMS OF A BACE INHIBITOR, COMPOSITIONS, AND THEIR USE	
9.	WO-2016053828-A1	2016-04-07	C07D 513/14	MERCK SHARP & DOHME
			C5-C6-FUSED TRICYCLIC IMINOTHIADIAZINE DIOXIDE COMPOUNDS AS BACE INHIBITORS, COMPOSITIONS, AND THEIR USE	
10.	WO-2016044120-A1	2016-03-24	C07D 471/04	MERCK SHARP & DOHME
			DIAZINE-FUSED AMIDINE COMPOUNDS AS BACE INHIBITORS, COMPOSITIONS, AND THEIR USE	

Figure 12. Results for a search in SureChEMBL with the selections specified in fig.11. Just the top-10 are shown from a total of 369 WO document hits.

Useful circumstantial information on SAR-likelihood can be gleaned from inspecting the retrieved title list even before opening up the individual documents (although less obvious aspects need some experience to be discerned). For example, we can see all three synonyms have been used by applicants and the patent titles appear to be true positives (n.b. we did not need this for the BindingDB search since they were pre-selected). It is also clear that we might have used "inhibitor" as another filter but this could have an associated false-negative risk. As we could already see from BindingDB, this larger result set confirms that BACE1 is a popular target since no less than five pharmaceutical companies are slugging it out even in just these 2016 publications. Another clue we can pick up on is that Merck have a secondary filing as "crystalline forms" which, by implication, should connect the structure back to an earlier novel active series that might include SAR. As additional context, we can also establish that both Merck and AstraZeneca have candidate BACE1 inhibitors still in clinical trials with Roche having had RG7129 in Phase 1 but terminated. We can thus infer that these recent filings from these three companies could include SAR from back-up and/or follow-on series for each of those clinical candidates. Roche have assigned identical titles to two different WO numbers that, by definition, should therefore include different chemistry claims (see IC50 search below).

Having assessed the titles we can zero in on some of these documents to illustrate particular points. The trio of titles from Lundbeck seem worth investigating because, as mentioned in the introduction, they could be a set of "D"s that have A-R-P in common and where "C" is an extended series. This turns out to be the case, with the three filings including individual Ki data tables for 41, 14 and 19 structures. We can pick out the following aspects from this set:

- They present a complex set of patent families but further down in the hit list we can find a yet a forth SAR set of 48 structures with Ki results in WO2015124576 from a 2-amino-3, 5, 5-trifluoro-3, 4, 5, 6-tetrahydropyridine series.

- SureChEMBL did a good job on all four sets. Notably, two had already appeared as US publications (e.g. WO2015124576 as US9353084 and WO2016075063 as US9346797) which, as expected, gave better extraction results.

- It turns out that WO2016075064 has unusual content. Firstly, there is a 3D crystal structure image of what was extracted as CID 121334268 . This provides a configuration for the series but had no reported activity data. Secondly, examples 16 and 18 were deuterated. This confounds a standard CNER pipeline both because the i2s did not recognise "D-" and the text OCR missed the "-d3" (although the n2s would have converted it within an IUPAC string) . To complete the SAR set a chemical sketcher could be used to input the two deuterated compounds (n.b. it may also happen that when this filing eventually surfaces at the USPTO the deuterated images will be linked to the document as molfiles via CWU processing).

Tools for stand-alone extraction

There are several options to collate SAR from US9346797 but this needs the example numbers to be matched to the data table on page 30 of the PDF. Analogous to the conversions of the example list indicated in figure 10, the sets are visible in SureChEMBL but in this case the applicants have usefully numbered them to match the Ki values that the document processing has extracted as text. While the SureChEMBL document chemistry export (at default filter settings) produced 114 rows that reduced to 55 unique structures, there was no easy way to isolate the 14 examples in the correct sequence. However, we can exploit the ordering by converting them extrinsically using a stand-alone operation. The resource of choice here was the open ChemAxon chemicalize resource that allows users to run n2s on text from a variety of sources an input formats [17]. In this case an external web page was made for batch conversion with just the 14 IUPACs pasted across from the SureChEMBL text (n.b. this may seem redundant since they have already been converted to structures but note that isolating just the selected strings and maintaining the order is important in this case). The results are shown in figure 13.

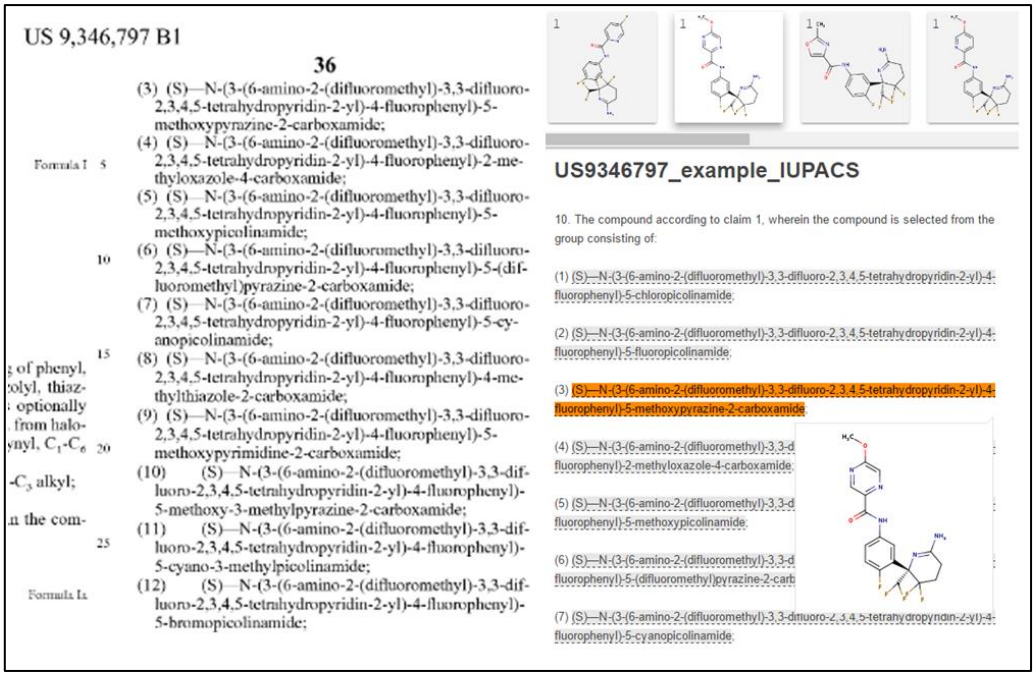
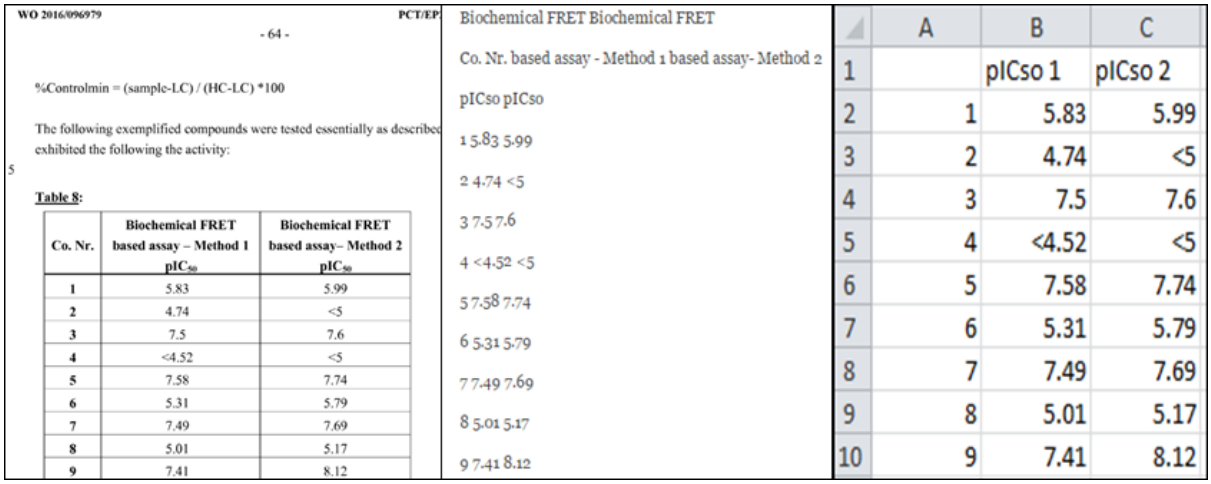


Figure 13. The SAR series from US9346797 extracted using chemicalize. Page 36 from the patent PDF specifying the structure is shown on the left. On the right are the chemicalize webpage conversions of the same numbered examples pasted over in the same order. The IUPAC for example three is highlighted in orange which also displays the ChemAxon structure rendering on mouse-over.

One of the features of chemicalize is that we can download an SD file from the webpage extractions in which the extraction order is maintained. This can then be easily transformed using commercial packages or open resources such as OpenBabel [18]. Thus, a combination of using SureChEMBL, chemicalize.org together with some manual collation and cross-checking against the original PDFs to tie things together, we can generate a complete Ki SAR set for the 96 tetrahydropyridines. Cases where SAR extraction presents more of challenge include the most recent in the figure 12 set , WO2016096979 from the Jansen team in Spain. We can quickly find SAR in the form of two different assays already log transformed to pIC50s cross-referenced against 33 compound numbers. Here again the OCR does a sufficiently good job on the data table to be copied over to Excel with the rows staying in register and the greater than signs as we can see in figure 14 (even though the out-of-range results might be rejected for model building).



WO 2016/096979 PCT/EP

- 64 -

%Controlmin = (sample-LC) / (HC-LC) *100

The following exemplified compounds were tested essentially as describec exhibited the following the activity:

Table 8:

Co. Nr.	Biochemical FRET based assay – Method 1 pIC_{50}	Biochemical FRET based assay– Method 2 pIC_{50}
1	5.83	5.99
2	4.74	<5
3	7.5	7.6
4	<4.52	<5
5	7.58	7.74
6	5.31	5.79
7	7.49	7.69
8	5.01	5.17
9	7.41	8.12

Biochemical FRET Biochemical FRET

Co. Nr. based assay - Method 1 based assay- Method 2

pICso pICso

1 5.83 5.99

2 4.74 <5

3 7.5 7.6

4 <4.52 <5

5 7.58 7.74

6 5.31 5.79

7 7.49 7.69

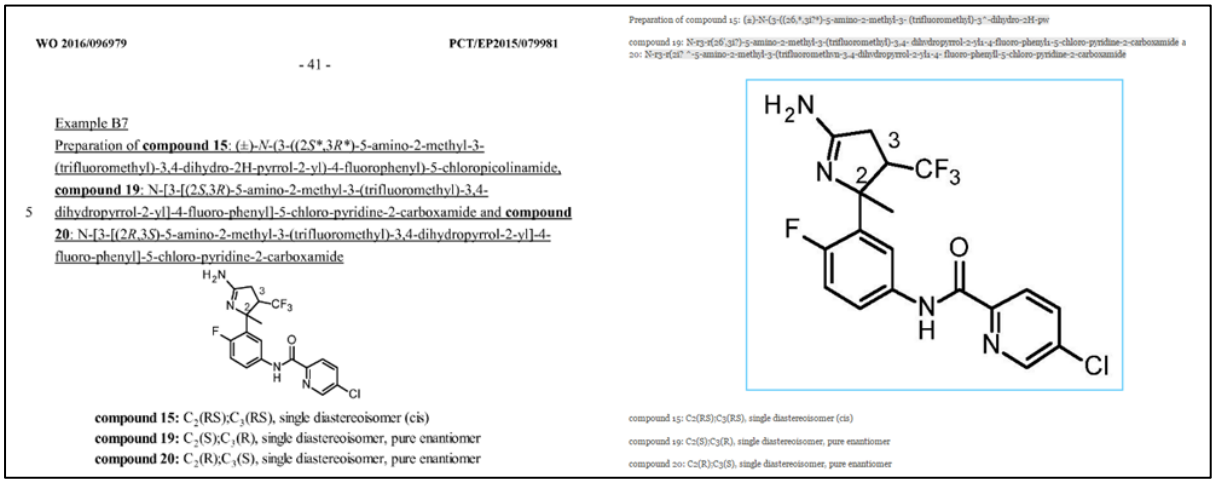
8 5.01 5.17

9 7.41 8.12

	A	B	C
1		pICso 1	pICso 2
2	1	5.83	5.99
3	2	4.74	<5
4	3	7.5	7.6
5	4	<4.52	<5
6	5	7.58	7.74
7	6	5.31	5.79
8	7	7.49	7.69
9	8	5.01	5.17
10	9	7.41	8.12

Figure 14. Three versions of the SAR table from WO2016096979. On the left is the original from page 64 of the PDF. In the centre is the corresponding section of the SureChEMBL mark-up. The right hand panel is an Excel paste-across of the centre section.

However, the failure of many of the ChEMBL chemistry conversions indicated that, compared to what we have looked at so far, it would not be straightforward to resolve compounds to match the numbers in the data table (figure 15).



WO 2016/096979 PCT/EP2015/079981

- 41 -

Example B7

Preparation of **compound 15**: (±)-*N*-(3-((2*S**,3*R**)-5-amino-2-methyl-3-(trifluoromethyl)-3,4-dihydro-2H-pyrrol-2-yl)-4-fluorophenyl)-5-chloropicolinamide, **compound 19**: N-[3-[(2*S*,3*R*)-5-amino-2-methyl-3-(trifluoromethyl)-3,4-dihydropyrrol-2-yl]-4-fluoro-phenyl]-5-chloro-pyridine-2-carboxamide and **compound 20**: N-[3-[(2*R*,3*S*)-5-amino-2-methyl-3-(trifluoromethyl)-3,4-dihydropyrrol-2-yl]-4-fluoro-phenyl]-5-chloro-pyridine-2-carboxamide

compound 15: C_2(RS);C_3(RS), single diastereoisomer (cis)
compound 19: C_2(S);C_3(R), single diastereoisomer, pure enantiomer
compound 20: C_2(R);C_3(S), single diastereoisomer, pure enantiomer

Figure 15. Chemistry for one example in WO2016096979. The compounds associated with Example B7 in the original PDF are shown on the left with the corresponding SureChEMBL section on the left. Note the "Highlight additional recognised chemical terms" option was toggled on. The consequent appearance of the grey highlighting indicates where the n2s recognised a potential name (i.e. an IUPAC-like text string) but could not generate a structure (in which case the highlighting would have been blue).

Inspection of the SureChEMBL output establishes that text quality had confounded the n2s algorithms. The extraction problems in this document were compounded by the i2s introducing an artefactual iodine from converting the B7 image (unfortunately now permanently enshrined in CID 121433102). One option would be wait for the US patent to surface with improved conversions. However, we can see the original document contains the necessary information we need for SAR collation but arranged in a somewhat complex way (e.g. example numbers are nested intermediates for the compound numbers that are also

depicted in Markush tables). Below is the SureChEMBL text associate with compound 19 from fig. 15, where the obvious OCR errors are marked in red.

N-**r3-r**(2**6'**,3i?)-5-amino-2-methyl-3-(trifluoromethyl)-3,4- dih**v**dropyrrol-2-yl**1**-4-fluoro-phenyl1-5-chloro-pyridine-2-carboxamide.

The tool we can use for stand-alone n2s is the Open Parser for Systematic IUPAC Nomenclature (OPSIN) [19] . While the web interface is only suitable for small numbers of conversions it has the advantage over Chemicalize of presenting error notifications and also allowing trial-and-error iterative editing of the input text. By this means we can correct compound 19 string to one that converts successfully

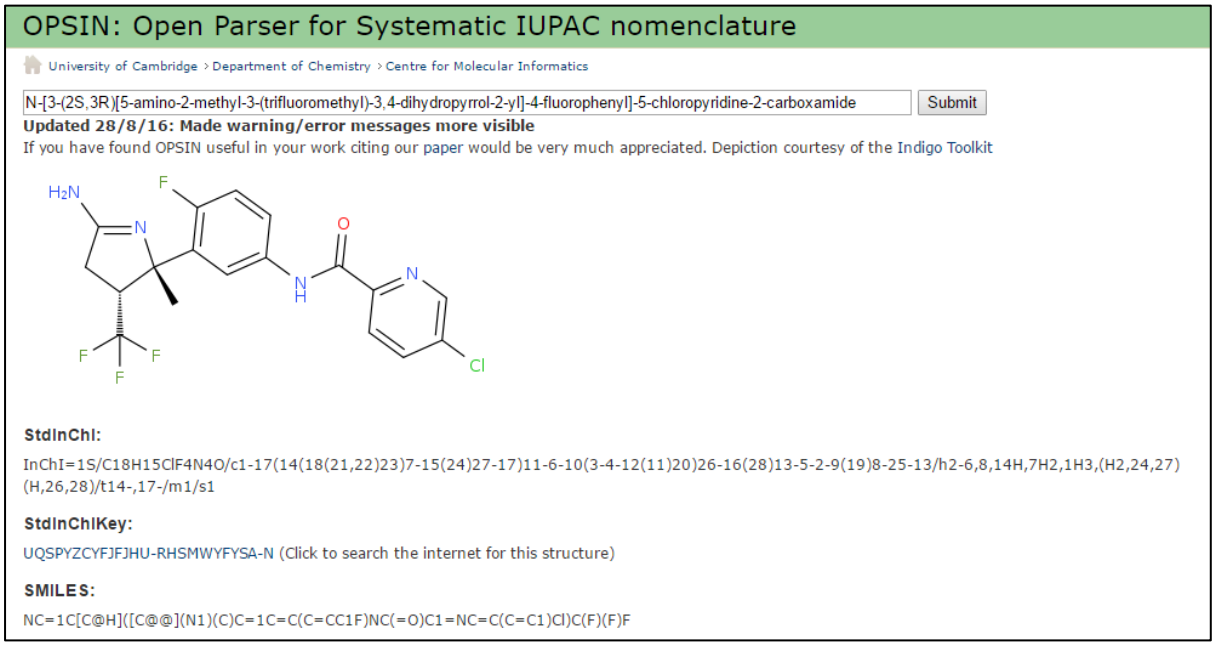


Figure 16. Conversion of compound 19 from WO2016096979 using the OPSIN interface. The input (after editing corrections) was N-[3-(2S,3R) [5-amino-2-methyl-3-(trifluoromethyl)-3,4-dihydropyrrol-2-yl]-4-fluorophenyl]-5-chloropyridine-2-carboxamide.

With some editing it now becomes possible to generate compound 15 as the “flat” chloropicolinoamide as well as specifying the resolved enantiomers of the carboxamide for 19 and 20 by editing the IUPAC S/R prefixes. To demonstrate another useful open tool set we can corroborate the i2s from the left hand side of figure 15 by using the Optical Structure Recognition Application (OSRA).

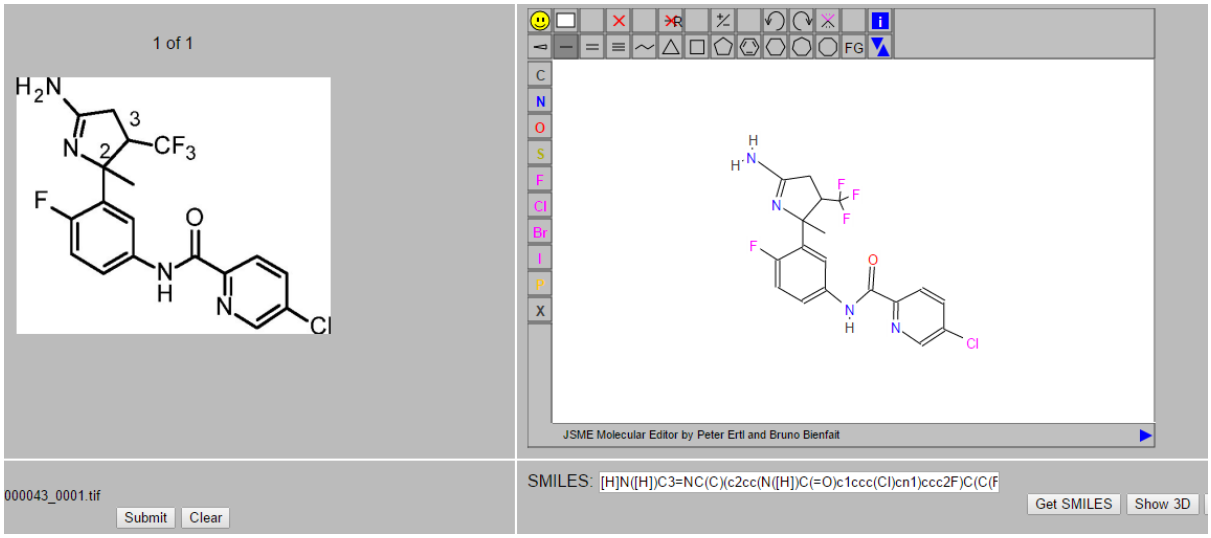


Figure 17. The OSRA result for conversion of the downloaded image from fig.15. The SMILES output can be converted to the example 15 IUPAC of N-{3-[5-amino-2-methyl-3-(trifluoromethyl)-3,4-dihydro-2H-pyrrol-2-yl]-4-fluorophenyl}-5-chloropyridine-2-carboxamide.

This WO2016096979 example shows the synergies of working between the original document, the SureChEMBL output with addition of the OPSIN and OSRA tools (chemcalize could also have utility here depending on individual preferences). While it has to said that collating the complete SAR set from these 33 examples, including the enantiomeric splits, would require some effort, the combined approach brings many other such "difficult" documents within practical reach of extractability. It is important to note that, due to partial failure of the SureChEMBL conversions, many of these examples will not be in PubChem.

Selecting for SAR

A particular advantage of full text indexing in resources such as SureChEMBL is that, regardless of the expected specificity issues, simple searches such as "IC50" or "inhibitor" actually work (108,115 and 650,147 documents respectively, with an AND union of 56,870). We can see a practical example in figure 18.

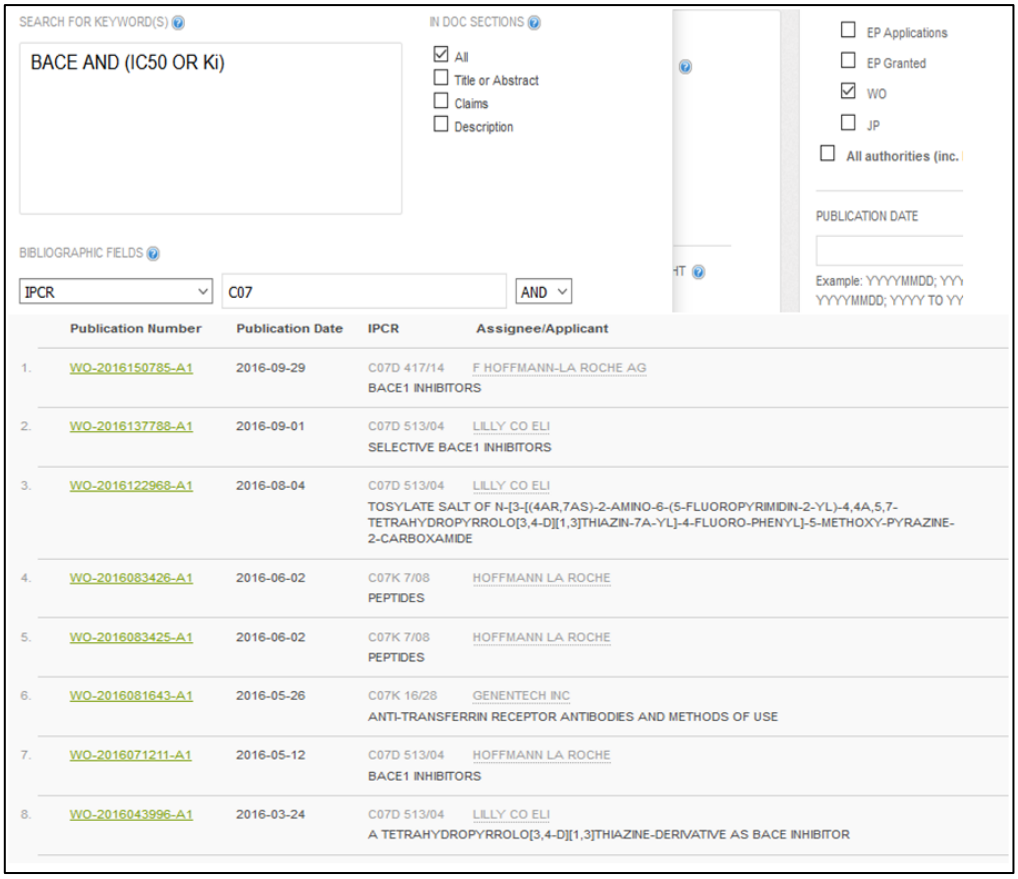


Figure 18. The most recent matches (from 138 in total) SureChEMBL search with for matches to BACE1 and IC50 or Ki and C07 within WO documents

We can see that the fig. 18 has matches in common with fig 12 and spot what seems a likely consecutive series from Roche as WO2016150785 and four months earlier WO2016071211 both titled “BACE1 inhibitors”. Inspection reveals these to be true-positives with WO2016150785 having seven intra-document IC50 matches including “Table 1 : IC50 values of selected examples”. While there are only four intra-document matches in WO2016071211 and the result table is not numbered, we can pick out both tables. This Roche appear to be true-positives for the IC50 trawl on the one hand, but may present borderline SAR-value on the other. The first issue is data sparcity (compared with what we have seen so far) since in WO2016071211 only 12 of the 25 IC50 examples have data and in WO2016150785 its 38 from 44. We can also see signs of probable deliberate obfuscation in the use of confusing nested numbering systems. However, a more serious issue we can immediately spot is difficulties for chemistry extraction because the poor OCR has confounded most of the n2s (including extensive introduction of the notorious 1-to-l conversion errors). While the i2s has fared better in at least getting structures out, the multiple conversions from small images in the tables present a confusing picture. We can look at an example in fig. 19.

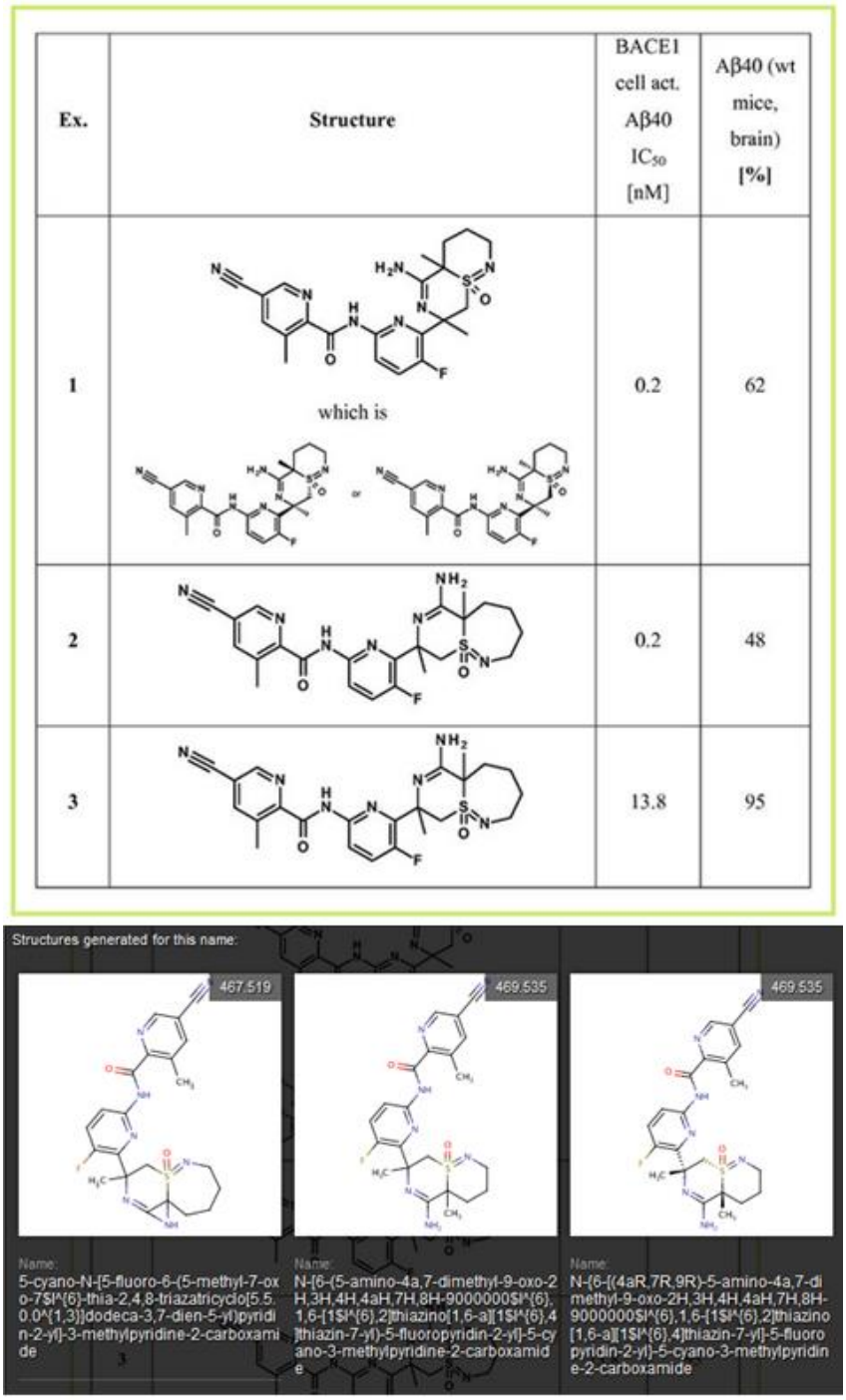


Figure 19. The first three examples from Table 1 in WO2016071211 are in the top panel. The SureChEMBL rendering is shown, with the green boarder signifying an intra-document search match. The image conversions from this section of the table have spawned the three structures in the lower panel (but not necessarily in order). From left to right in the panel these are in PubChem as CID 121329828, CID 121329766 and CID 121329765.

We can follow these structures "back round" in the sense of mapping them back to their locations in the document (via PubChem or inside SureChEMBL). However, we see a peculiar pattern of stereo enumerations and "flat" structures that are matching multiple different example numbers (e.g. two and three are identical). The minimum parsimonious assumption is that these multiple mappings could be due to errors in the document. At the

end of the day this example can be seen as a cautionary tale. Thus, WO2016071211 would seem unreliable for SAR modelling and not worth spending time with any tools to try to sort things out. However, thanks to the SureChEMBL "efforts" we can at least see the series chemotype.

Connecting between patents and papers via PubChem

Explorations of patent SAR naturally generate a range of questions with respect to the literature. The two basic ones are (Q1) "have any of the bioactive structures in this patent appeared in papers? The converse question (Q2) is "have any of the bioactive structures in this paper (in particular the lead compound) appeared within a (possibly larger) SAR series in a patent ? These questions thus look forwards or backwards in time and can both be important. There are technically different ways to approach the questions and with different search stringencies (e.g. an exact match, a connectivity match, a similarity match or multiple matches across a series). Before looking at some examples it is important to understand the time relationship of what surfaces where when, how much and who cross-points to whom. Starting with patent chemistry in PubChem first we can establish that it contains 16.3 million substances from SureChEMBL up to the 6th of August. We can also check that *in situ* the latest extractions were only three days behind the publication date. While the lag-time in PubChem is thus about a month, this is more than acceptable for a free resource. There are four other sources of patent chemistry in PubChem that each have more than a million structures (see table 1 in [2] for a 2015 snapshot) but three of these are static and are not likely to update. Selecting SureChEMBL, IBM, SCRIPDB and NextMoveSoftware (all CNER sources) adds up to 19.7 million compounds with 82% coming from SureChEMBL.

PubChem is extensively connected to the literature via different types of CID-to-PubMed connectivity [20] one of which includes ChEMBL via the PubChem BioAssay database. However there is an associated lag time for when PubMed IDs (PMIDs) are connected to CIDs, because ChEMBL release times are well over six months. Nonetheless, there are multiple ways to answer Q1 and Q2. One of these is by following the reciprocal cross-pointers between PubChem, SureChEMBL, ChEMBL and UniChem. This means for any individual structure we can follow inter-database and paper-patent connectivity without doing any searches at all (this was already demonstrated for SureChEMBL-to-PubChem). Before we explore search examples we should examine the general question of what is the overall overlap between structures in papers and patents. The commercial GOSTAR figure is that 9% of compounds from papers are found in their patent extractions, indicating a slight rise from 6% in 2009 [6]. By contrast 18% of ChEMBL has a SureChEMBL match. There are two possible reasons for the higher number. Firstly, ChEMBL subsumes ~0.5 million compounds from confirmed Molecular Libraries Screening Centers Network (MLSCN) assays so this substantial increase in structures would be expected to increase in turn the number of literature extraction matches. The second is that SureChEMBL extracts approximately three times the number of compounds GOSTAR does, which should also increase the % overlap with literature extraction.

We can approach the questions by looking at Q2. A simple PubMed search with "BACE1 inhibitor" brings up a relevant looking paper high in the result list (fig. 20).

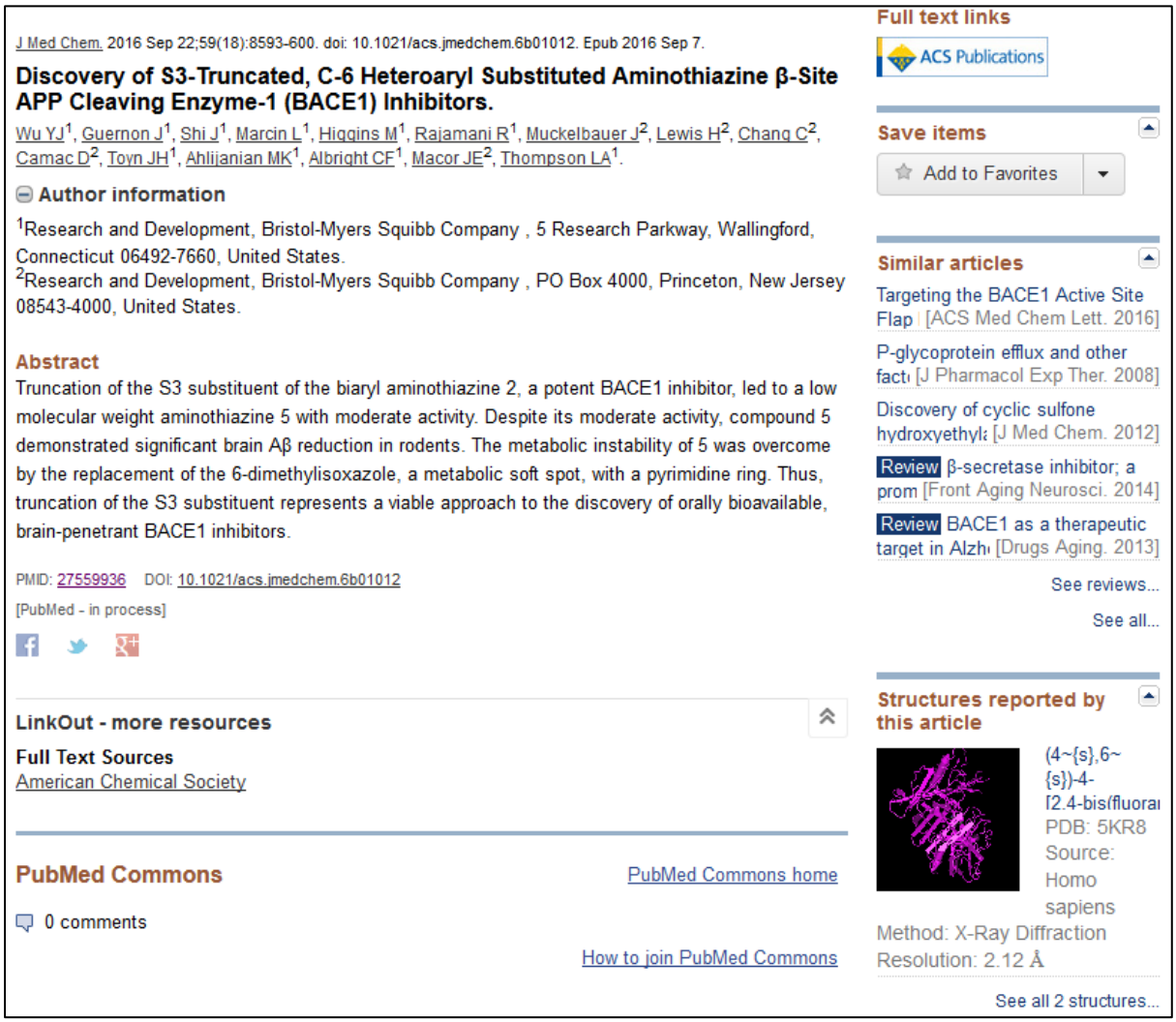


Figure 20. The 4th ranking match in PubMed with the query "BACE1 inhibitor" .

This very recent paper (not yet captured by ChEMBL) from Bristol-Meyers (BMS) has the hallmarks of a drug discovery team pursuing an orally available, brain penetrant BACE1 inhibitor for which they already have activity in a mouse model. The bonus for our question answering is that they have deposited two PDB structures containing ligands from the paper. Because the NCBI MMDB system assign a CID to PDB ligands where possible the resulting CID 90253397 connects us right through to the SureChEMBL patents (fig. 21).

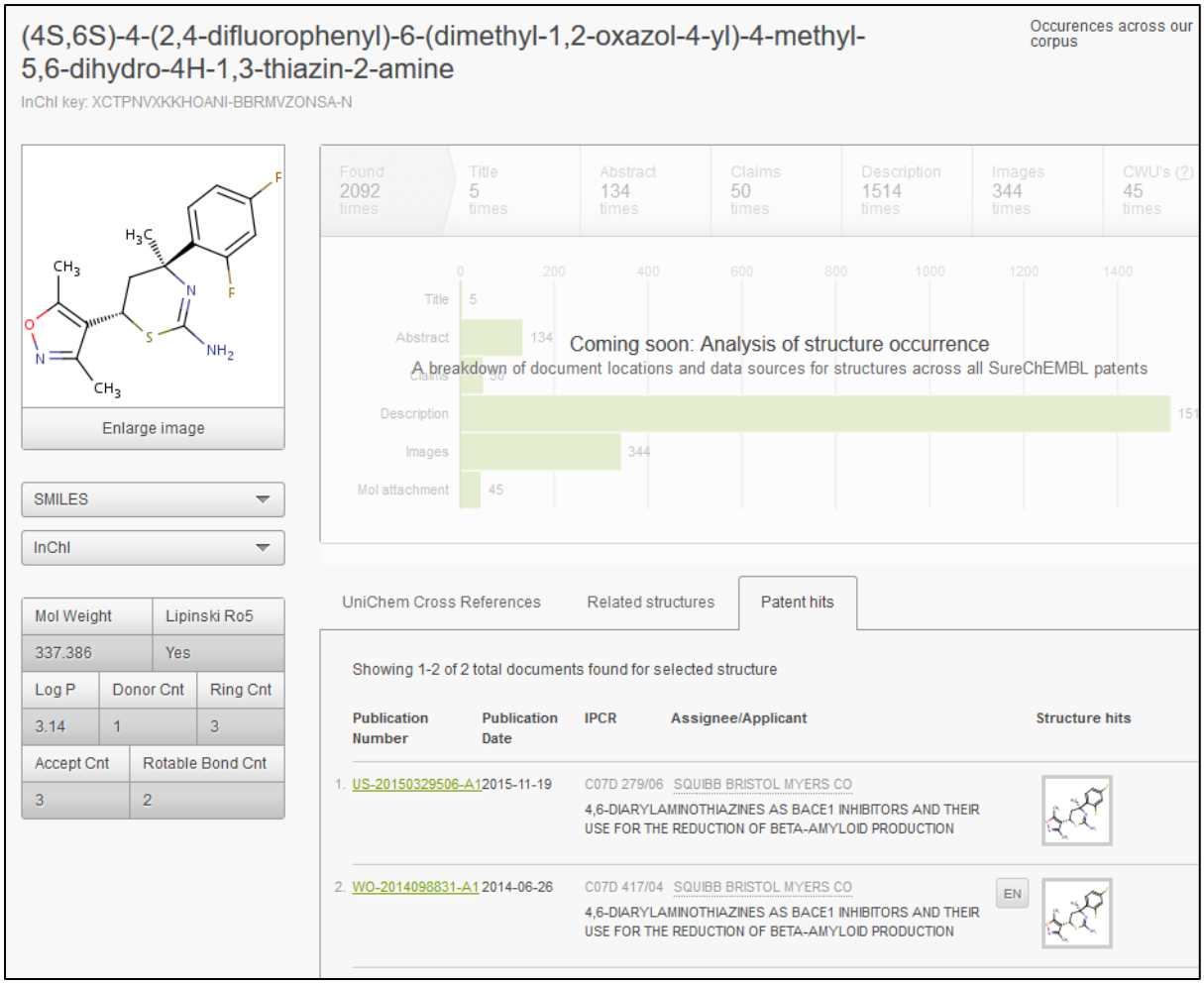


Figure 21. The connection between the publication PMID 27559936, the PDB entry 5KR8, the CID 90253397, the SureChEMBL SID 240769892 and the two family BMS patents above.

While the BMS paper has good SAR characterisation for 18 structures in total [21], the patent has cell based IC50 data for a series of 62 compounds that, with some manual engagement, can all be identified from the SureChEMBL output. Note also that WO2014098831 was picked up in the fig. 12 search but was a year further down the list. Of the possible aproaches to Q1 perhaps the most effective is to download filtered patent chemistry from SureChEMBL and upload it to PubChem using the Indentifier Exchange Service. The next step is to intersect these CIDs with ChEMBL. Lead-like matches (i.e. not common chemistry) are likely to be from journal extractions at least two years or more after the patent date.

Tracking back to first-filings

This is a variation on the paper-to-patent Q2 but here the focus is searching backwards for the early patents on clinical candidates or recently approved drugs. The special interest here is to possibly access patent SAR around a compound that has reached some level of success, even just in the preclinical stages. One resource that enables making the appropriate connections is the IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) [22] . Inspection of the

BACE1 entry indicates the expert curation of 15 lead compound and clinical candidates as ligand entries. For some of these patents have been curated into the reference where these were found to contain useful SAR. A section from this panel is shown in fig. 22 and the SureChEMBL page shown in fig.23.

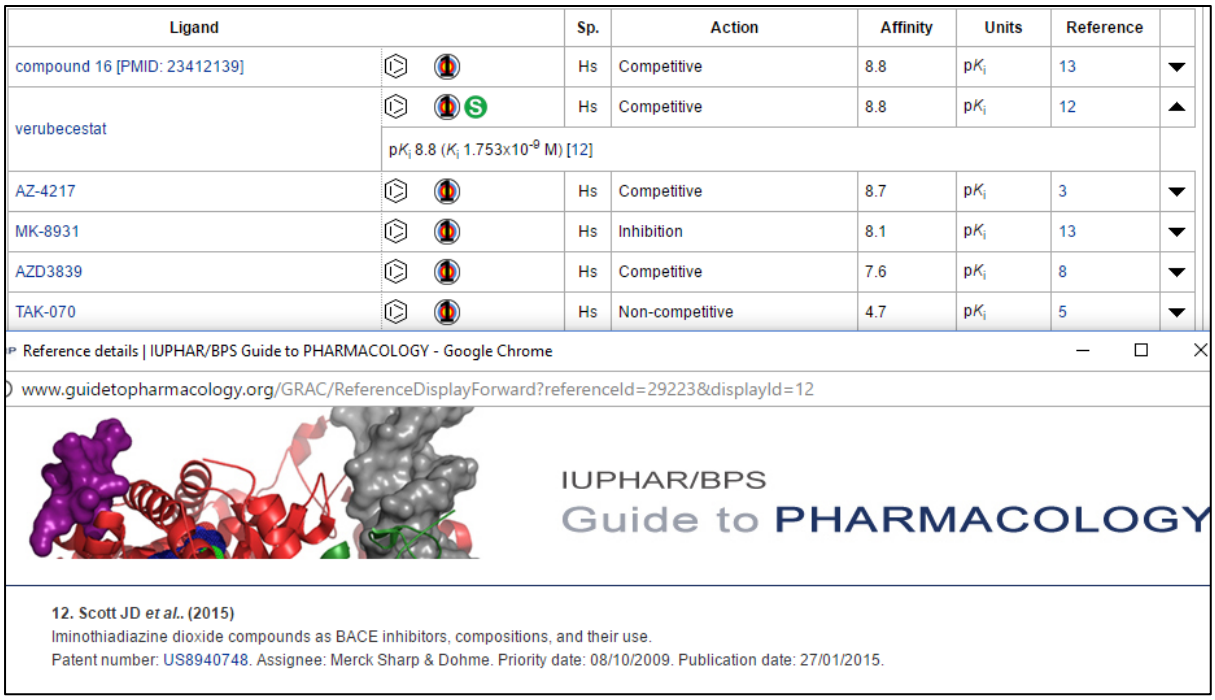


Figure 22. Snapshot from the GtoPdb page for BACE1 ligands. The record for verubecestat is expanded to show the pKi, the patent reference in the lower panel, a live link to the SureChEMBL patent document and identification of the example number as 25.

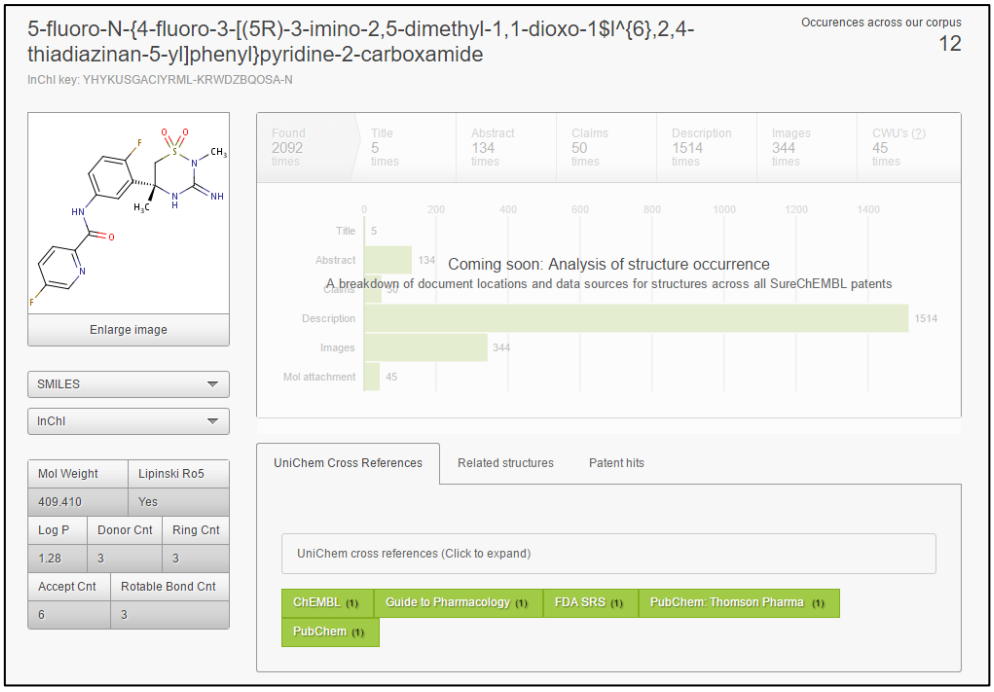


Figure 23. The SureChEMBL page for verubecestat. The lower set of green panels show the cross pointers to other databases mediated via UniChem, including a GtoPdb link to fig. 22.

Note also if the 14 patent document hits are opened up we see Merck WO2016053767 for "Novel crystalline forms of a bace inhibitor, compositions, and their use" which was the secondary patent pointed out in fig.12.

Given the verubecestat compound from Merck is now in Phase III it is somewhat unusual in not having any journal publications. This, in this case the curation of a patent link into GtoPdb is particularly useful. Although the discrete the SAR in US8940748 is minimal there are BACE1 and BACE2 Ki values. Also somewhat unusually there are in the order of 200 analogue structures full described without inhibition data but many with permeability measurements. N Using the approaches described above most lead compounds and drugs, using GtoPdb or other resources as entry points, can now be tracked back to their first-filings as novel structures, many of which include SAR.

Conclusions and prospects

It is hoped that the approaches and examples presented here will not only encourage but also enable medicinal chemists to dig SAR out of patents. The recent "big bang" of patent chemistry flowing into public databases, alluded to above, now presents the best of both worlds situation in two senses. This first is that the open patent data is becoming more synergistic with the features and content of commercial resources. The second is that those without access to the latter are no longer constrained for SAR-digging even it may take a bit more time and effort. So what of the future? It would be useful to see comparative modelling results that might illuminate similarities and differences between data sources (i.e. could adding patent-only SAR improve models?)

Other developments will include significant optimisation of the PubChem patent system in 2017 (Evan Bolton, personal communication). In addition, the development of patent information extraction systems continues (e.g. the 2015 BioCreative V challenge "Chemical and Drug Named Entity Recognition from patent text" CHEMDNER patents [23]). As an extension of these efforts both the commercial and open sectors are pursuing the holy grail of full automated D-A-R-C-P extraction using NLP and deep indexing approaches. In conjunction we may also see "big data" scale patent processing initiatives in the public domain. These can exploit Application Programming Interfaces (APIs) web services (WS) and Resource Description Frameworks (RDF). These are already becoming available for SureChEMBL as part of the OpenPHACTS initiative [24]. So might the major patent offices move towards fully electronic patent applications that would include XML full text, direct submission of chemical structures, use of bioassay ontologies, standardised data tables and bioentity identifiers for C07/A61 applications? We'll see

Note added in proof: As an important new development, in October 2016 the WIPO PATENTSCOPE portal (see website below) released a chemical structure search feature for ~ 7 million compounds they have already extracted from WO documents.

References

[1] Southan C, Boppana K, Jagarlapudi SAA, Muresan S. Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds. J Cheminform 2011;3:14+. doi:10.1186/1758-2946-3-14.

[2] Southan C. Expanding opportunities for mining bioactive chemistry from patents. Drug Discov Today Technol 2015;14:3–9. doi:10.1016/j.ddtec.2014.12.001.

[3] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res 2016;44:D1202-13. doi:10.1093/nar/gkv951.

[4] Southan C, Hancock JM. A tale of two drug targets: the evolutionary history of BACE1 and BACE2. Front Genet 2013;4:293. doi:10.3389/fgene.2013.00293.

[5] Clark AM, Dole K, Ekins S. Open Source Bayesian Models. 3. Composite Models for Prediction of Binned Responses. J Chem Inf Model 2016;56:275–85. doi:10.1021/acs.jcim.5b00555.

[6] SOUTHAN C*., Várkonyi P, Muresan S, Várkonyi P, Varkonyi P, Southan. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. J Cheminform 2009;1:10. doi:10.1186/1758-2946-1-10.

[7] Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, et al. Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discov Today 2011;16:1019–30. doi:10.1016/j.drudis.2011.10.005.

[8] Eriksson M, Nilsson I, Kogej T, Southan C, Johansson M, Tyrchan C, et al. SARConnect: A Tool to Interrogate the Connectivity Between Proteins, Chemical Structures and Activity Data. Mol Inf 2012;31:555–68. doi:10.1002/minf.201200030.

[9] Ekins, Sean, Clark, Alex M., Southan, Christopher, Bunin Barry A, Williams AJ. Small-molecule Bioactivity Databases. In: High Throughput Screening Methods: Evolution and Refinement, editor. Nathan Ross Joshua Bittker, Royal Society of Chemistry, in press; 2016, p. 344–65.

[10] Southan C, Varkonyi P, Boppana K, Jagarlapudi SARP, Muresan S. Tracking 20 years of compound-to-target output from literature and patents. PLoS One 2013;8:e77142. doi:10.1371/journal.pone.0077142.

[11] Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 2016;44:D1045-53. doi:10.1093/nar/gkv1072.

[12] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014;42:D1083-90. doi:10.1093/nar/gkt1031.

[13] Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, et al. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res 2016;44:D1220-8. doi:10.1093/nar/gkv1253.

[14] Southan C. BACE2 as a new diabetes target: a patent review (2010 - 2012). Expert Opin Ther Pat 2013;23:649–63. doi:10.1517/13543776.2013.780032.

[15] Chambers J, Davies M, Gaulton A, Papadatos G, Hersey A, Overington JP. UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. J Cheminform 2014;6:43. doi:10.1186/s13321-014-0043-5.

[16] Schneider N, Lowe DM, Sayle RA, Tarselli MA, Landrum GA. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. J Med Chem 2016;59:4385–402. doi:10.1021/acs.jmedchem.6b00153.

[17] Southan C, Stracz A. Extracting and connecting chemical structures from text sources using chemicalize.org. J Cheminform 2013;5:20. doi:10.1186/1758-2946-5-20.

[18] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform 2011;3:33. doi:10.1186/1758-2946-3-33.

[19] Lowe DM, Corbett PT, Murray-Rust P, Glen RC. Chemical name to structure: OPSIN, an open source solution. J Chem Inf Model 2011;51:739–53. doi:10.1021/ci100384d.

[20] Kim S, Thiessen PA, Cheng T, Yu B, Shoemaker BA, Wang J, et al. Literature information in PubChem: associations between PubChem records and scientific articles. J Cheminform 2016;8:32. doi:10.1186/s13321-016-0142-6.

[21] Wu Y-J, Guernon J, Shi J, Marcin L, Higgins M, Rajamani R, et al. Discovery of S3-Truncated, C-6 Heteroaryl Substituted Aminothiazine β-Site APP Cleaving Enzyme-1 (BACE1) Inhibitors. J Med Chem 2016;59:8593–600. doi:10.1021/acs.jmedchem.6b01012.

[22] Southan C, Sharman JL, Benson HE, Faccenda E, Pawson AJ, Alexander SPH, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res 2016;44:D1054-68. doi:10.1093/nar/gkv1037.

[23] Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BFH, Hettne KM, et al. Chemical entity recognition in patents by combining dictionary-based and statistical approaches. Database (Oxford) 2016;2016. doi:10.1093/database/baw061.

[24] Chichester C, Digles D, Siebes R, Loizou A, Groth P, Harland L. Drug discovery FAQs: workflows for answering multidomain drug discovery questions. Drug Discov Today 2015;20:399–405. doi:10.1016/j.drudis.2014.11.006.

Relevant Websites

SureChEMBL (large scale patent chemistry extraction)
https://www.surechembl.org/search/

PubChem (92 million structures)
https://pubchem.ncbi.nlm.nih.gov/

PubChem Identifier Exchange Service (mapping extrinsic chemistry into PubChem)
https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi

BindingDB (binding data for proteins and small molecules)
http://bindingdb.org/bind/index.jsp

ChEMBL (large scale literature extraction)
https://www.ebi.ac.uk/chembl/

UniChem (cross-referencing of 137million structures including ChEMBL, SureChEMBL and PubChem)
https://www.ebi.ac.uk/unichem/

OSRA: Optical Structure Recognition Application (i2s)
https://cactus.nci.nih.gov/osra/

OPSIN: Open Parser for Systematic IUPAC nomenclature (n2s)
http://opsin.ch.cam.ac.uk/

ChemAxon Chemicalize, text processing (n2s, with registration required)
https://chemicalize.com/welcome

IUPHAR/BPS Guide to PHARMACOLOGY (leads with patent connectivity)
http://www.guidetopharmacology.org/

GOSTAR (Excelra Knowledge Solutions)
http://www.excelra.com/gostar.php

WIPO PATENTSCOPE database of 58 million documents
http://www.wipo.int/patentscope/en/

European Patent Office (EPO) Espacenet database of 90 million documents
https://worldwide.espacenet.com/

USPTO Patent Full-Text and Image Database
https://www.uspto.gov/patents-application-process/search-patents

International Patent Classification (IPC)
http://www.wipo.int/classifications/ipc/en/